Computational Identification of Recessive Mutations in Cancers Using High Throughput SNP-arrays

Marko Laakso

Helsinki January 10, 2007

M. Sc. Thesis

UNIVERSITY OF HELSINKI
Department of Computer Science

Acknowledgments

This Master's thesis in science has been written for Department of Computer Science at University of Helsinki. The colorectal study has been conducted in Tumor Genomics research group and I have been analysing the data in Computational Systems Biology Laboratory. Both research groups belong to Faculty of Medicine in University of Helsinki.

I am grateful because Professor Juho Rousu volunteered to supervise this thesis. He has taken care that the content of this thesis meets the requirements of Department of Computer Science. I would like to thank Juho for pointing out the blurry and misleading expressions from my text and for his contribution to Master's degree programme in bioinformatics.

I would have never made this project without Docent Sampsa Hautaniemi, who hired me into his laboratory. Sampsa has given me a unique chance to combine my two favourite duties, programming and biological research. With his guidance I have acquainted myself with some of the wonders of scientific community. In his post, as a leader of the laboratory, he has shown excellent skills in combining the interests of employees, science and the society. There is a lot I have learned from our discussions!

Working in Computational Systems Biology Laboratory has been a rousing experience. Biomedicum provides a wide range of lectures and events about various aspects of medical research and we have been blessed to collaborate with many top-ranked research groups. I am proud of my co-workers: Anna-Maria, Jianmin, Kari, Odenna, Sirkku, and Susanna as they express skills in so many different sciences.

The biomedical part of the thesis is based on the work that has been done in Professor Lauri Aaltonen's laboratory. The detection of mutant candidate regions and their scoring has been developed together with Lauri Aaltonen, Auli Karhu, Rainer Lehtonen and Sari Tuupanen. These people have done great work in teaching me cancer genetics. I want to praise Auli's and Rainer's supervision on this thesis. Weekly meetings and many other events in Aaltonen's lab have given me some insight to cancer research and wet lab procedures. I would like to thank Anniina, Annika, Antti, Heli L., Heli S., Iina N., Iina V., Inga-Lill, Laura, Maija, Marianna, Matti, Mia, Mikko, Pia A, Pia V., Päivi H., Päivi L., Rainer, Reijo, Sakari, Sanna, Silva, Sini, Susa, Taru, Tuija, Ulla-Maija, and Virpi for the opportunity to watch their work.

The sample material used in this study has been collected from 42 colorectal cancer patients and 51 siblings of leukaemia patients, all being more than aware of the suffering caused by the cancer. I wish we can make the most of your donations and perhaps all

this helps people in future. I am grateful for The Finnish Red Cross (especially to Hannu Toivonen and Jukka Partanen) for the reference data they provided. Getting that many controls would have been very hard without your help. Janna Saarela (National Public Health Institute) analysed the reference samples and she also helped in delivering the data to us. Our CRC project was financially supported by Academy of Finland, Biocentrum Helsinki, Finnish Cancer Organisations and the Sigrid Jusélius Foundation, which are gratefully acknowledged.

Scientific Computing Ltd. (CSC) has provided us computing resources for the haplotype analysis and CohortComparator experiments. The CSC facilities have been working reliably and we have always got top quality assistance as needed. Yrjö Leino's preliminary results of the code optimisations have been very impressive. It has been a real honour to get such a professional to help us.

Talking Glossary of Genetic Terms [Nat06] provided by National Human Genome Research Institute has been a good source of illustrations. The DNA and chromosome drawings of Figure 1 are based of their original images.

Helsinki, January 2007
Marko Laakso

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta/Osasto — Fakultet/Sektion — Faculty	Laitos — Institution — Department
Faculty of Science	Department of Computer Science

Tekijä — Författare — Author
Marko Laakso

Työn nimi — Arbetets titel — Title
Computational Identification of Recessive Mutations in Cancers Using High Throughput SNP-arrays

Oppiaine — Läroämne — Subject
Bioinformatics

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
M. Sc. Thesis	January 10, 2007	63 + 9 pages

Tiivistelmä — Referat — Abstract

This thesis presents a highly sensitive genome wide search method for recessive mutations. The method is suitable for distantly related samples that are divided into phenotype positives and negatives. High througput genotype arrays are used to identify and compare homozygous regions between the cohorts. The method is demonstrated by comparing colorectal cancer patients against unaffected references. The objective is to find homozygous regions and alleles that are more common in cancer patients.

We have designed and implemented software tools to automate the data analysis from genotypes to lists of candidate genes and to their properties. The programs have been designed in respect to a pipeline architecture that allows their integration to other programs such as biological databases and copy number analysis tools. The integration of the tools is crucial as the genome wide analysis of the cohort differences produces many candidate regions not related to the studied phenotype.

CohortComparator is a genotype comparison tool that detects homozygous regions and compares their loci and allele constitutions between two sets of samples. The data is visualised in chromosome specific graphs illustrating the homozygous regions and alleles of each sample. The genomic regions that may harbour recessive mutations are emphasised with different colours and a scoring scheme is given for these regions.

The detection of homozygous regions, cohort comparisons and result annotations are all subjected to presumptions many of which have been parameterized in our programs. The effect of these parameters and the suitable scope of the methods have been evaluated. Samples with different resolutions can be balanced with the genotype estimates of their haplotypes and they can be used within the same study.

ACM Computing Classification System (CCS):
G.3 [Probability and statistics],
J.3 [Life and medical sciences]

Avainsanat — Nyckelord — Keywords
bioinformatics, homozygosity, SNP

Säilytyspaikka — Förvaringsställe — Where deposited
Kumpula Science Library, serial number C-

Muita tietoja — övriga uppgifter — Additional information

Contents

Abbreviations and Symbols **vi**

1 Introduction **1**

2 Cancer Genetic Background **4**

2.1 Introduction to cancer biology 4

2.2 Mutations and markers 6

2.3 Detection of cancer related mutations 8

3 High Throughput SNP-microarray Analysis **10**

3.1 SNP oligonucleotide array 10

3.1.1 Composition and usage of SNP-microarrays 11

3.1.2 Affymetrix SNP-microarrays 12

3.2 Data preprocessing 13

4 Analysis Methods for Homozygous Regions **15**

4.1 Definition of the homozygous region 15

4.2 Detection Algorithms 18

4.3 Scoring 22

4.4 Compound heterozygosity 23

5 Annotations **27**

6 Materials and Methods **30**

6.1 Origin of the data 30

6.2 Programs 31

6.2.1 AffyExportParser 33

6.2.2 fastPHASE 34

6.2.3 CohortComparator 35

6.2.4 RegionAnnotator 37

7 Results **39**

7.1 Distribution of homozygous SNPs 39

7.2 Determination of a significant region length 40

7.3 Properties of the region detection algorithms 43
7.4 Mutation detection probability 48
7.5 Effects of the haplotyping 49
7.6 Biological content of the candidate regions 52

8 Discussion **55**

References **58**

Appendices

1 Graphs of Interesting Regions

2 Parameters of CohortComparator

Abbreviations and Symbols

$\underline{\text{aa}}$	homozygous combination of alleles A and A
$\underline{\text{ab}}$	heterozygous combination of alleles A and B
$\underline{\text{bb}}$	homozygous combination of alleles B and B
bp	base pair
$\mathbf{C}^{\{F\}}$	region matrix of compound heterozygotes (F=filtered)
CGH	comparative genomic hybridisation
CRC	colorectal cancer
CSC	Center for scientific computing (CSC - Scientific Computing Limited)
d	the minimum number of homozygous SNPs required between two heterozygous ones
DNA	deoxyribonucleic acid
g	maximal rate of heterozygosity
GO	Gene ontology
$h^{\{D,R\}}$	probability of having a homozygous region at a certain region (D=patient data, R=reference data)
HMM	hidden Markov model
k	number of measured SNP loci
$l^{\{D,R\}}$	minimum length of regions accepted (D=patient data, R=reference data)
LOH	loss of heterozygosity
Mbp	mega base pairs = 1 000 000 bp
MM	mismatch probe
n	number of samples
PM	perfect match probe
$\mathbf{R}^{\{l,gd,sw,D,R\}}$	region matrix containing 1 for homozygous SNP and 0 for heterozygous SNP (l=filtered with l, gd=gap distance model, sw=sliding window model, D=patient data, R=reference data)
MIAME	Minimum information about a microarray experiment
RNA	ribonucleic acid
$\mathbf{S}^{\{D,R\}}$	sample matrix of SNP haplotypes (D=patient data, R=reference data)
s	maximal length of a heterozygous gap within a homozygous region
SNP	single nucleotide polymorphism
$t^{\{D,R\}}$	number of overlapping homozygous regions (D=patient data, R=reference data)
w	length of the sliding window given in SNPs

1 Introduction

Cancers are life threatening diseases that are becoming more common as people live longer [AD04]. There are several factors affecting the development of a cancer. For example some chemicals, such as tobacco, asbestos, radioactive substances, and benzene are well known for their cancer inducing properties but it is also known that environmental factors such as sunlight and other ionizing radiation may also contribute to the development of the cancer. Yet another well known fact is that some viral infections may lead to cancer [Wei06]. A common property of all cancer causing agents is their ability to alter the cellular DNA content. The alteration may be caused by either a direct damage to the molecule itself or it can be a stress mediated process in which other cellular processes are disturbed decreasing the ability of the cell to maintain its genomic integrity.

The genetic origin of cancers and the variation between the individual genomes partly explains why some people are more subjected to certain types of cancers. It has been shown in many studies that some families are carrying mutations increasing their risk of cancer [Knu71, Sal00, HP04, Vie06, Web06]. The awareness of such mutation helps in screening the risk groups and to treat them in a more precise manner. The identification of mutations and the affected genes is essential to achieve new diagnostics and therapies.

The detection of cancer related mutations provides a challenging task for bioinformatics as the human genome contains over 21000 know genes [Bir06]. High throughput genotyping methods such as single nucleotide polymorphism oligonucleotide arrays described in Section 3.1 can be used for the genome wide analysis of the DNA samples. The data processing tools described in this thesis can be used to process the vast amount of the data generated by the new measurement techniques.

The amount of publicly available genotype data increases as many journals[1] presume the release of original measurements for the articles referring to such data. Minimum Information About a Microarray Experiment (MIAME) [Bra01] defines a guideline stating what kind of background information should be provided along the microarray measurements to make them reproducible. The information consists of experiment details such as sample and probe descriptions that are essential for the analysis of the data. The international HapMap [The03] project collects information about genotype frequencies in human populations and the data is publicly available on their Internet site. HapMap information can be used to select genetic markers suitable for the population under concern.

This thesis focuses on the process of refining the SNP haplotype data into something with

[1] For example: Bioinformatics, Nature, and PLoS Genetics

more biological relevance such as scored and annotated lists of regions with significant difference to controls. The differentiated region of samples will in turn guide further experimentation and sequencing of the prospective DNA segment. The implementation of a SNP data set visualisation and comparison tool play an important role in this thesis. The key thing is to find a sensitive method to detect homozygous regions and to compare them between the two sets of data. The interest is not only in those homozygous regions that are heterozygous in controls but we are also keen in locations with different allele[2] compositions or significant overlapping homozygous deletions. It was found in the preliminary phase of the project that straightforward statistical analyses with exact p-values are not efficient means to address the sensitivity requirements. A scoring schematic with gene annotations was found to be a better alternative. The major objective of the project is to find potential regions harbouring inherited mutations in cancer predisposing genes. Based on results emerging from developed algorithm experimentally focused scientists can then choose the best candidates for the wet lab verifications.

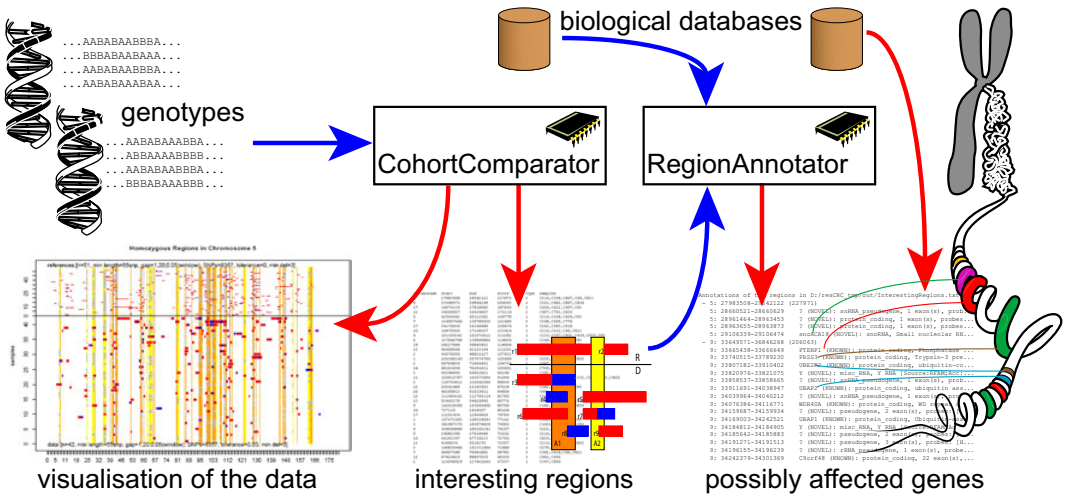

Figure 1: An overview of the data inputs (blue arrows) and outputs (red arrows). Genotype information of two sample sets is visualised in chromosome specific figures. The homozygous regions and their differences in sample sets are listed into text files. RegionAnnotator assigns gene annotations for each interesting region. The gene list can be further coupled with other biological databases or literature.

The analysis of genotype data is based on two sets of data. The samples are divided into these two categories based on the *phenotype* of individuals that is their observable or measurable traits. The detection of interesting genomic differences between the sample sets is based on the assumption that the phenotype may be caused by a recessive mutation

[2] An allele refers to a certain nucleotide at a certain locus

in some samples. Only homozygous deletions are analysed due the unrelated nature of samples and because the analysis has been conducted using the genotype data. Methods requiring parent-offspring trios [Con06] are clearly unsuitable but the integration of chromosomal copy numbers may improve our results as explained in Section 8. A phenotype criterion of colorectal cancer has been used in this study but the comparison methods are presumably applicable to other phenotypes also.

2 Cancer Genetic Background

A short introduction to the necessary biological concepts in colorectal cancer is given in this section. The means of genetic variation are explained as they affect the analysis but the detailed mechanisms of inheritance and gene regulations are not explained in this paper. Good sources of additional cancer information are for example [Wei06] and [VK02].

2.1 Introduction to cancer biology

Cancer in this thesis is defined as malignant neoplasia, which is a genetic disorder in cells. Cancer cells are dividing in an uncontrolled manner and invading to new tissues. The invasion may concern the neighbouring tissues or some more distant parts of the body in which case the invasion is referred as *metastasis*. Approximately 90% of cancer victims are killed by the metastases making them far more dangerous than the primary tumours [HW00].

Cancer cells are not only multiplying in an uncontrolled manner but they can also induce vascular growth, escape immune system and avoid apoptosis [HW00]. *Apoptosis* is a process of programmed cell death that is usually triggered when the cell is either subjected to overwhelming stress such as virus infection or DNA damage or detected by the immune system [KF00]. The transformation from a normal cell to a cancer is usually due the accumulation of mutations leading to all required properties [Fea98, p. 229]. The amount of initial mutations in cancers varies typically from two to about ten but they tend to accumulate during the progression of the disease and the number of mutated genes in mature tumours reaches an average of about 90 [Sjö06]. Germ cells and their predecessors, so called germ line cells, may pass their mutations to subsequent generations. Cells, other than those in a germ line are called *somatic cells*. The first mutations may have been inherited from the parents as *a hereditary mutation* and the second mutation may occur in somatic cells [Knu71, MK06]. It is possibly that mutations have occurred in somatic cells, in which case the cancer is *sporadic*. *Loss of heterozygosity* (LOH) occurs when the originally heterozygous tumour cells become homozygous for a locus and the allelic contribution of the other parent is lost. The second mutation is usually a LOH that can be caused by a deletion in homologous chromosome or by a mutation that removes the difference between alleles.

The concept of *gene* has become ambiguous as a wide variety of cellular DNA func-

tions have been discovered [Pea06]. Here, gene is used to refer transcripts (that are RNA products of expressed genes) of any kind. All genetic variations affecting the cancer susceptibility are of our interest and thus no distinctions are made between protein coding genes and other transcripts like microRNAs [Gar06]. The sequence annotations are based on transcripts available in Ensembl database as described in Section 5.

The most dramatic cancer causing mutations are usually affecting oncogenes, DNA repair genes, or tumour suppressor genes. *Oncogenes* are cancer-inducing genes [Wei06, p. G:14] and they are usually up-regulated in cancer cells. The function of oncogenes is typically related to control of the cell differentiation and proliferation. Changes in *DNA repair genes* may decrease their activity and promote further alterations in genome thus increasing the chance of developing a cancer. Typical, *tumour suppressor genes* encode proteins that regulate the growth of the cell. The function of tumour suppressor genes is complementary to oncogenes in a sense that they inhibit processes responsible for the cancerous behaviour of the tumour cell. Inactivation of the tumour suppressor genes may promote the cancer genesis. Phenotypic changes in tumour suppressor genes require a loss of both copies in a region of the gene since the inactivation of the gene in one chromosome is usually compensated by the unaffected gene in the homologous chromosome.

Colorectal cancer (CRC) is the third common type of cancers [Ame06, Fin05]. There were about 2300 new cases diagnosed and 1100 persons died because of CRCs in Finland during the year 2003 [Fin05]. The risk of getting a CRC is doubled if it has been diagnosed from a first degree relative and even four times the average if the relative was younger than 45 years old at the time of the diagnosis [JH01]. Those who have inherited a cancer predisposing mutation are more likely to develop cancer and they may do so while younger because they already have the initial mutation. Mutant carrying cells need less (and may be more prone to) somatic alterations in order to become malignant. Several genes are known to increase CRC risk if they are mutated (for example *APC*, *MYH*, *SMAD4*, *ALK3*, *STK11/LKB1* and DNA mismatch repair genes) [Web06]. Still, a fraction of the families of CRC patients may harbour mutations in genes that are not known to be cancer predisposing [Aal]. The effects of an ancient recessive mutation can be seen in those who have inherited the mutation from their both parents that are descendants of the person who had the original mutation.

Isolated populations with extensive and reliable genealogical records are ideal for genetic studies of hereditary diseases [PJV99, VP04, MK06]. Suitable populations can be found for example from Quebec, Iceland, Northern Sweden, and Finland [VP04]. The chance of consanguineous parents increases in these populations and thus recessively heritable phe-

notypes are seen more frequently. The genealogical records can be used to map certain genotypes to mutant carrying individuals if the affected people are close relatives and the records are available. The methodologies of these linkage studies are further discussed in Section 2.3. In this study we have concentrated on developing an autozygosity mapping method suitable for surveys of distantly related individuals. We assume an isolated population in order to attain a reasonable chance of autozygosity while the number of required samples is kept in minimum.

2.2 Mutations and markers

Cellular DNA is packed into highly coiled molecules called *chromosomes*. *Homo sapiens* is a *diploid* species since it has two copies of each homologous chromosome. *Homologous chromosomes* are corresponding chromosomes with a common origin and they have separate copies of the same genes. The length of the chromosomes varies between 47 and 247 Mbp having total of over 3433 Mbp [Bir06]. The whole genome consists of 23 pairs of chromosomes [3]. All but one (1–22) of these pairs are autosomes and the last pair (23) consists of sex chromosomes. *Autosomes* are chromosomes that are not responsible for the gender of an individual. The normal diploid constitution of sex chromosomes is either a pair of two X chromosomes (female) or one X and one Y (male).

Recombination of the sections of chromosomes takes place during the production of germ cells (meiosis). The regions of chromosomes that are inherited in continuous pieces are called *haplotypes*. The original haplotype of the person who got the mutation degrades to ever smaller fractions as the generations go by and it gets mixed with the other haplotypes. The various aspects of haplotypes are further discussed in Section 4.4.

The mutation itself creates some variation to the genome of the population but it is also surrounded by a vast amount of other mutations that have accumulated during the evolution in the context of its haplotype. The frequency of mutations to be inherited together is called *linkage* and two alleles are said to be in *linkage disequilibrium* if their linkage is greater than what would be expected for the independent alleles. The variation in DNA may consist of deletions where some pieces are missing, insertions of new genetic material or changes in nucleotides, where a sequence is changed to another. Most of the variation in human genome consists of substitutions in single nucleotide, where one of the four nucleotides (adenosine, thymine, guanine, and cytosine) has changed to another

[3] To be precise, each mitochondrion has its own copy of an additional chromosome but these extranuclear chromosomes are ignored in this thesis

one. The phenomenon of having such a varying nucleotide at a certain locus is referred as *single nucleotide polymorphism* (SNP). Some of these alterations have been associated to certain phenotypes [The03]. Common definition of the SNP requires that the relative frequency of the least frequent allele is greater than 0.01 [Moo05]. The development of the high throughput methods of detecting these SNPs has given us a chance to trace haplotypes based on their SNP content. An example of a mutation haplotype is illustrated in Figure 2. Individuals A and B share a mutation between the SNP_2 and SNP_3. The phenotype effects of the mutation can be seen in both individuals if the mutation is dominant. Individual A expresses no recessive mutations as the mutation is not present in chromosome A_2. A recessive mutation can be restricted to the right side of SNP_1 since SNP_1 is heterozygous in B. The detection of the mutation requires a complete sequencing of the DNA between SNP_1 and a marker that follows SNP_3 but has not been associated to the same haplotype.

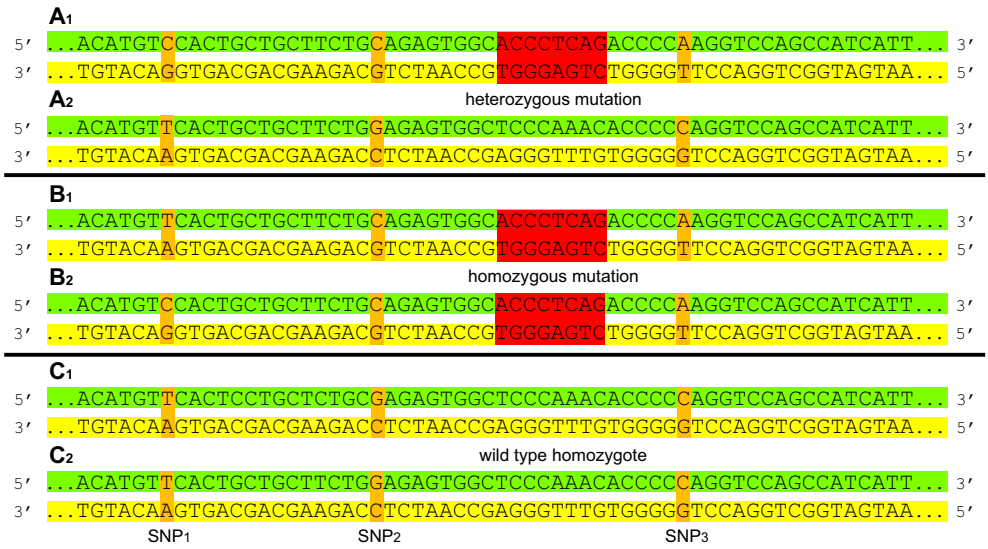

Figure 2: Illustration of the relationship between SNPs and haplotypes. Three SNPs measured from 3 individuals (A, B, and C) are listed with two strands of DNA (green and yellow) and two chromosomes (1 and 2). The varying alleles for the SNPs in 5'-strand are: $SNP_1 \in \{C, T\}$, $SNP_2 \in \{C, G\}$, and $SNP_3 \in \{A, C\}$. The mutation is associated to haplotype of $SNP_2 = C$ and $SNP_3 = A$.

Recessive phenotypes are expressed if a mutant allele is inherited from both parents. The chromosomal regions that have the same alleles in both homologous chromosomes are called homozygous (a more detailed definitions of the homozygous region are given in Section 4.1). Identification of a homozygous region from affected individuals suggests the existence of a causative mutation within the region [MK06]. A homozygous region

is called *autozygous* if both copies of the allele are inherited from the common ancestor. A chromosomal region is called *hemizygous* if the genome contains only one copy of an allele and its homologies. Hemizygous cells might have lost a chromosome or its fragments. Homozygous regions can be caused by a somatic LOH or as typically, because of autozygosity. We do not make difference between homozygous and hemizygous regions and thus we denote $\underline{aa}$ = {AA, A}, $\underline{bb}$ = {BB, B}.

The vast majority of the human genome is homozygous in a sense that the sequence is identical in both homologous chromosomes. The similarity of the sequences is especially noted for genes and other evolutionary conserved parts of the genome. Even those loci that vary between individuals are likely to be homozygous as one of the alleles may be more common than the others. Distributions of two alleles can be estimated using the binomial distribution if a random mating is assumed within the (infinitely large) population as stated by Hardy-Weinberg principle [Ste43]. The genotype frequencies are $f(\underline{aa}) = f(\underline{a})^2$, $f(\underline{ab}) = 2f(\underline{a})f(\underline{b})$, and $f(\underline{bb}) = f(\underline{b})^2$, where $f()$ refers to the frequency of the allele or the genotype.

2.3 Detection of cancer related mutations

In this section we describe how cancer related mutations are studied in traditional cancer biology. First, we illustrate the concept of linkage studies and the use of pedigrees in them. Second, we describe how to find mutations using genotype data and smaller pedigrees. The distant relatedness of the patients and the paucity of pedigrees form a difference between the scopes of linkage and association studies such as the CRC study described in this thesis.

The detection of a cancer related mutation is usually motivated by the observation of several similar cancers in different individuals. Traditionally, relationships between the patients are used to trace possibly inherited mutations [Mid04, MK06, Vie06]. The pedigrees can be sorted out using the knowledge about the ancestors and offspring of the patients. Information about phenotypes is bind to pedigrees in order to identify lineages of mutant inheritance. Some ancestors of the affected individuals may be identified as carriers of the mutation under the hypothesis that the mutation exists. The number of the mutant carrying individuals that develop a cancer and gets diagnosed depends on the age of the individuals [AD04] and the penetrance of the mutation [HP04]. Most cancers are not caused by germ line mutations [Lic00]; especially they are not necessarily affected by the same mutations [Sjö06]. Some people may have developed a cancer due a series

of somatic mutations and they may be identified as separate family lineages or they may confound the analysis producing misleading patterns of inheritance.

Genome wide sequencing of the patient samples would be too laborious with the current techniques and thus some marker (microsatellites or SNPs) based heuristics [Abe02] are used to infer haplotypes that may contain the mutation. First, haplotype patterns of markers in linkage disequilibrium are constructed. Secondly, haplotype patterns are compared against the pedigrees to identify patterns that would explain the observed phenotypes.

Haplotypes that match the patterns of affected individuals can be used to identify chromosome regions with possible mutations. The biological content of the region can be fetched from biological databases as explained in Section 5. Functional experiments can be conducted to see whether some of the genes found from the regions are differentially expressed in cancer patients in comparison to unaffected individuals. The sequences of differentially expressed genes should be compared in order to identify the actual genetic difference between the haplotypes.

A good example of the described cancer gene detection methods can be found from pituitary adenoma study [Vie06], where a mutation of *AIP* gene was found from people in Northern Finland. The affected people were found to be members of two family lineages suggesting a germ line mutation.

The linkage based detection of mutations with a low penetrance requires a vast number of samples from affected siblings, which can be attenuated by the methods used in association studies [HP04]. The association methods can be used for the identification of polygenic susceptibilities as the risk of cancer can be posed by a set of alleles [Web06]. None of the alleles is necessarily highly penetrative but their combination may cause a significant increase in the cancer predisposition [HP04]. Relative frequencies of haplotypes between case and control samples can be compared statistically if there are enough samples [Onk02]. The haplotypes, which are more frequent in case samples may contribute the observed phenotype.

Recessively inherited mutations should be identifiable in autozygous regions of those expressing the phenotype [LB87]. The autozygous regions can be identified from the genotype sequences produced by the SNP-microarray experiments but unfortunately current solutions assume close relatedness of the samples [GT02, Woo04, Chi06]. In this thesis, we proposed a new procedure for unrelated samples. The procedure scales to high density array data sets of hundreds of samples. The method takes genotyping errors into account and it can be adjusted to non-uniform marker distributions.

3 High Throughput SNP-microarray Analysis

Single nucleotide polymorphisms are one of the most abundant genetic variations among humans [Moo05]. The vast number of SNPs causes a challenge for biologists and bioinformaticians although they provide lot information about the relationships between individuals. Genomic variation is traditionally analysed using microsatellite markers [WCK06]. *Microsatellites* are short, highly repetitive DNA sequences and their length varies among the individuals. The SNPs are not as informative markers as longer sequences but the vast density and potential for automation makes them useful in genetic studies [Wan98]. The appearance of new microarray methods has provided a cost-efficient way to analyse over 10^6 SNPs from one sample. The haplotype information provides insight into evolutional backgrounds of the populations. HapMap [The03] is an extensive international project collecting information about the genetic variation among the humans. The population based information about the allele frequencies can be used to trace origins of the differences.

The SNP-microarrays are used for example in chromosomal copy number analysis, linkage analysis [Mid04], LOH detection [Ber06, LT00, Lin04, ZPH05], and in general genotyping. An SNP-microarray experiment produces an estimate of the allele concentrations for each SNP. The concentrations are relative to each other (see normalization in Section 3.2). The genotype analysis is based on a comparison between the expected alleles, where equal strength of the signals of both alleles indicates heterozygosity and the hybridisation of one allele is considered a homozygous SNP. The copy number analysis is based on more accurate comparison of the concentrations to predict the multiples of alleles [Nan05, Tin06, Zha04, Zha05]. Tools and resources available for the SNP marker studies are reviewed in [Moo05], which gives a more detailed overview of biological potential of SNPs.

3.1 SNP oligonucleotide array

The use of microarrays has been in a rapid expansion in biological and medical research ever since their invention in late 1980's. SNP-microarrays provide a relatively fast and inexpensive genotyping method, which requires only tiny amounts of DNA. Different kinds of technologies are used in SNP-microarray production and platforms vary in their hybridisation and staining techniques [Aff04, Jaa03, Syv05]. Commercial arrays are marketed with brands such as: Asper, TaqMan, Sequenom, Illumina, and Affymetrix. The overall concept of DNA sequence specific hybridisation between a set of reference se-

quences and the sample sequences is, nevertheless, conserved.

3.1.1 Composition and usage of SNP-microarrays

SNP-microarrays are small chips covered by an array of tiny droplets of short single-stranded DNA fragments called *probes*. These fragments of some 25 to 50 nucleotides are complementary to sequences around the SNP loci. During the hybridisation, some fragmented and stained sample DNA is put on the array. The single-stranded sample fragments binds to probes with a complementary sequence. The longer oligonucleotides are more sensitive and they work with lower DNA concentrations. The shorter sequences are, however, more specific in binding to their complements. To date, there is no consensus in selecting the optimal length for the oligonucleotides. The protocol used in sample hybridisation on SNP chip depends on the platform in use but the general concept is to pour a sample of labelled and fragmented DNA on the chip. The sample fragments are let to hybridise with the matching complements attached to the chip. The excess sample material is washed away and the chip surface is ready to be scanned to a bitmap image.

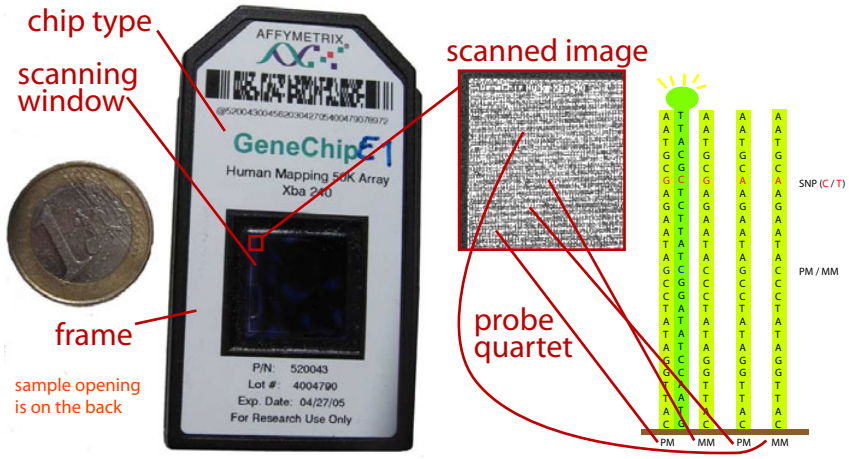

Figure 3: Size and structure of the XbaI chip. An imagined example of a single probe quartet is shown on the right. The quartet consists of perfect match and mismatch probes of both alleles. The distinction between the PM and MM is always made in the middle of the sequence, whereas the SNP locus varies between the quartets. An example target sequence with a labelling dye residue has bound to the left most probe, which represents allele C. The other PM probe has not bind to its target and the sample is considered to be homozygous for the C allele.

SNP-microarrays can be produced in biological laboratories equipped with a microarray printing facilities. The SNP density of the custom made chips is usually relatively small comparing to industrial chips but they can be made for such organisms and for such genomic regions that are not covered by the industrial chips available. The creation of a

SNP-microarray is a laborious and difficult process not due to challenges to select probes for the array. The probe selection requires a good knowledge or plausible assumptions about the genetic variation among individuals so that the SNP loci can be identified and the whole genome must have been sequenced in order to make sure that the probes are unique to locus of interest. A single stranded DNA molecule may fold so that it binds to itself if it contains complementary sequences of its own. The probe selection involves the verification that there will not be problems with the folding of the molecule. The target sequence of the probe should be within a range of DNA that can be amplified in order to establish high DNA concentrations for the hybridisation.

3.1.2 Affymetrix SNP-microarrays

Affymetrix is a company that is specialised in the production of complementary DNA microarrays. The same photolithography based manufacturing procedure is used for the production of gene expression and SNP-microarrays, and the physical properties of the chips are similar. The distinction between these two array types is in the origin of the sample DNA. The whole genome is used in SNP analysis whereas the expression chips expect complement DNA of the message RNA molecules restricted from the cells. The photolithographic synthesis of the probes is done by extending the probes of the selected spots with one nucleotide at a time [Aff06]. The mount of the chip is first coated with a light-sensitive chemical that prevents the nucleotides from binding it. The initial coating is followed by a series of steps.

1. The probes are selected to be extended with a certain nucleotide by using a mask that allows UV-light to remove their light-sensitive protection.
2. The chip is next flooded with a solution of single nucleotides of one type. Each nucleotide carries a light-sensitive group protecting it against binding to other nucleotides. The nucleotides are let to bind to the unprotected ends of the selected probes.
3. The excess of nucleotides is washed away and the process is repeated with another nucleotide.

Affymetrix provides a wide variety of different microarrays for different purposes and for different organisms. Genome wide SNP-microarrays for humans are currently available at resolutions of about 10000 SNPs, 2×50000 SNPs, and 2×250000 SNPs. The combinations of 2 chips work so that the sample DNA is divided into two sets that are digested

with different restriction enzymes producing fragments based on different splicing sequences for PCR amplification. The joint utilisation of the restriction enzymes and the PCR amplification of the sequences up to 2000 bp allow the reduction the complexity of the genome by producing copies of regions targeted by the chips [Syv05]. Affymetrix GeneChip®Human Mapping 100K Set utilises two chips with restriction enzymes XbaI and HindIII. Both chips are capable of measuring over 50000 SNP loci consisting of over 2500000 features that is different kinds of probe sequences. Each SNP is targeted with 40 different probe sequences consisting of ten different probe quartets. A probe quartet consists of a mismatch (MM) and perfect match (PM) probes of both target nucleotides of the SNP. Perfect match probes match the sequence around the SNP loci and mismatch probes having a wrong nucleotide in the middle of the probe. The chip architecture assumes that there are only two varying nucleotide candidates for each SNP locus. The use of MM probes and misclassification of unexpected alleles is discussed in Section 3.2. An illustration of a fictional probe quartet is given in Figure 3. The exclusive hybridisation (green sample fragment with a fluorescent biotin label) of the left most perfect match probe indicates homozygosity for *C* allele (no hybridisation to other PM probe) and good separation between the probes (no cross hybridisation to MM probes).

3.2 Data preprocessing

The data processing starts from the scanned image the first step being the detection of the hybridisation spots and their intensities. The sequence fragments from the sample can be detected as fluorescent spots on the array since they carried the labelling dye to their hybridisation position. The exact sequence is known as the underlying complement sequence. The intensity of the spot tells us about the abundance of the sequence indicating whether the sequence is present or not. Comparing the abundance information between consecutive SNPs and the values of a control sample with a known copy number gives an opportunity to estimate the chromosomal copy number [Nan05, Tin06, Zha04, Zha05].

Preprocessing of the SNP-microarray data covers the process of detecting haplotypes of individuals from the bitmap image produced by the scanner. The standard preprocessing process consists of five major steps: grid detection, intensity calculations of the spots, quality diagnostics, normalization, and estimation of the haplotypes.

Grid detection involves the recognition of the array of rows and columns of different oligonucleotides. The fitted grid is usually an array of rectangles with a certain amount of tolerance for irregularities of the array and the produced image.

Intensities of the spots are the actual information provided by a SNP-microarray experiment. Determination of the intensity is a nontrivial task subjected to various sources of errors such as irregularities of the shape of the spot, spatial variation in scanner sensitivity, levels of hybridisation, and unexpected intensity distribution within a grid bounded area.

The intensities of the spot should be of the same scale if they are to be compared against each other. The process of the value scaling and shifting is called normalization. One of the major reasons for the normalization is to reduce the non-biological variation that is noise in data. A typical SNP-microarray contains various spots that measure the same region of the genome and it is assumed that their values should correlate in a manner that depends on their relation to the SNP, which is usually same value for the probes of the same allele. The assumption behind normalization is that the overall distribution of the intensities is known and the measurements follow it.

The purpose of the mismatch probes of Affymetrix GeneChip®Human Mapping chips is to measure the level of non-specific hybridisation that is considered to be the background intensity. The intensities of the PMs are determined based on their difference in corresponding MM intensities. The constitution of alleles in each SNP can be determined using the dynamic modelling (DM) algorithm that compares the likelihoods of models of aa, ab, bb, and `NoCall` [Di05]. The algorithm makes no clear distinction between homozygous and hemizygous loci and the most likely model is used as a genotype estimate. The determination of the SNP alleles is made in a SNP and chip specific manner. The genotype estimates of the multi-chip studies may use the robust linear model with Mahalanobis distance classifier (RLMM) that utilises the information about the similarities between the probes and genotypes of different chips [RS05].

There are various causes for missing values in SNP genotypes [Car06]. The undetected alleles can be seen as `NoCall` values or hemizygous SNPs depending on whether they occur in one or both homologous chromosomes. Some of these undetected, so called *null alleles* can be caused by a deletion [Con06]. The deletions can be of biological origin but they may also indicate sample degradation or other difficulties in the SNP experiment. The samples with high rates ($>0.05k$) of `NoCall` SNPs are likely to be degraded or the experiment has failed [Aff04]. Some of the null alleles can be explained by another polymorphism so close to the SNP site that it interferes with the probes and there is also a chance of having more than two alleles within the population. Affymetrix arrays are built so that they are targeting only two known allele polymorphism but the structure of DNA allows total of 4 (adenine, thymine, cytosine and guanine) different nucleotides for each locus.

4 Analysis Methods for Homozygous Regions

Detection of homozygous regions from the normal tissue of distantly related samples is one of the most important goals in this study. A homozygous region is a chromosomal region with the same allele inherited from both parents. The aspects of this definition will be further discussed in Section 4.1. Section 4.2 describes algorithms that can be used to detect homozygous regions from SNP-microarray data. The sample-wise comparison of the regions is discussed in Sections 4.3 and 4.4, which describe the concepts of interesting region, region score and compound heterozygotes.

4.1 Definition of the homozygous region

The concept of a homozygous region refers to a genomic subsequence containing a continuum of homozygous loci. The consecutive SNP samples from such region should be homozygous and we can assume that a sequence of such homozygous SNPs contains homozygous nucleotides between the SNP loci. The exact base pair bounds of the homozygous regions cannot be given using plain SNP data as it gives no information about the nucleotides between the bounding homozygous and heterozygous SNPs. Here, we extend homozygous regions all the way to the next heterozygous SNP loci, which mean that we may have coupled some heterozygous nucleotides into consideration but at least we do not miss the homozygous ones.

Let there be n samples each consisting of k SNP loci. A trivial way of defining a homozygous regions matrix $\mathbf{R} \in \{0,1\}^{n \times k}$ for the samples $\mathbf{S} \in \{\underline{aa}, \underline{ab}, \underline{bb}\}^{n \times k}$ is to create a Boolean function indicating the homozygous loci of each sample as

$$\mathbf{R}_{r,c} = \begin{cases} 1, & \mathbf{S}_{r,c} = \underline{aa} \vee \mathbf{S}_{r,c} = \underline{bb} \\ 0, & \mathbf{S}_{r,c} = \underline{ab} \end{cases}, \qquad r \in [1,n], c \in [1,k]. \tag{1}$$

Now the Boolean matrix $\mathbf{R}$ contains series of ones representing the homozygous regions. Too short regions can be filtered out by selecting sequences of at least l ones in a row. Another way of saying this is that homozygous SNP at $\mathbf{R}_{r,c}$ is surrounded by $l-1$ other homozygous SNPs. Table 1 illustrates how we can calculate homozygous SNPs around the given (r,c) in all possible l ways. $\mathbf{R}_{r,c}$ belongs to a long enough sequence if and only if the sum of any of these counts becomes l. A formal representation for the length limit can be written as:

$$\mathbf{R}^l_{r,c} = \bigvee_{v=c-l+1}^{c} \left(\left(\sum_{i=max(v,1)}^{min(v+l,k)} \mathbf{R}_{r,i} \right) = l \right), \qquad l \in [1,k]. \tag{2}$$

A)		c	k	
$\mathbf{R}_r$	$\ldots 0\,0\,1\,0$	1	$1\,1$	$\sum\ l$
				$2 \neq 4 \rightarrow 0$
				$3 \neq 4 \rightarrow 0$
				$3 \neq 4 \rightarrow 0$
				$3 \neq 4 \rightarrow 0$
$\mathbf{R}_r^l$	$\ldots 0\,0\,0\,0$	0	$0\,0$	

B)		c	k	
$\mathbf{R}_r$	$\ldots 0\,0\,1\,1$	1	$1\,0$	$\sum\ l$
				$3 \neq 4 \rightarrow 0$
				$4 = 4 \rightarrow 1$
				$3 \neq 4 \rightarrow 0$
				$2 \neq 4 \rightarrow 0$
$\mathbf{R}_r^l$	$\ldots 0\,0\,1\,1$	1	$1\,0$	

Table 1: Illustration of the Equation 2. SNP (r, c) is part of a sequence of length 3 in case A and the $\mathbf{R}^l_{r,c}$ is 0 because $l = 4$. Case B shows how one of the summations becomes 4 leading to the recognition of the sequence.

The idea in Equation 2 is to test all possible l length sequences containing $\mathbf{R}_{r,c}$. The disjunction is satisfied if there is at least one sequence without a heterozygous gap. Index i is kept within [1,k] by using the min and max functions.

A difficulty in defining the concept of homozygous region arises when there are inaccuracies in the data and we may see some heterozygous SNPs in otherwise homozygous region. There are many different possibilities to handle the noise in data but the approach, that is chosen affects to results and thus it is important to understand the behaviour of the approach in use. All methods must favour regions with a higher rate of homozygous SNPs for homozygous regions and reject regions with a too low density in order to be useful.

A simple step towards non-uniform regions is to define the minimal distance d between any two gaps within a region. Each gap of heterozygous SNPs must be shorter than s. The formal definition of this *gap distance* model can be derived in two steps using Equation 2. First, a matrix $\mathbf{R}^{gd'}$ of homozygous fragments of at least d continuous SNPs is created:

$$\mathbf{R}^{gd'}_{r,c} = \bigvee_{v=c-d+1}^{c} \left(\left(\sum_{i=max(v,1)}^{min(v+d,k)} \mathbf{R}_{r,i} \right) = d \right), \qquad d \in [1, l]. \tag{3}$$

Second, those fragments in $\mathbf{R}^{gd'}$ that are separated by s or less SNPs are merged. The conjoining of the fragments and the application of the length limit l can be coupled into form of:

$$\mathbf{R}^{gd}_{r,c} = \bigvee_{j=c}^{c+l-1} \left(\bigwedge_{v=j-l-s+1}^{j+s-2} \left(\left(\sum_{i=max(v,1)}^{min(v+s,k)} \mathbf{R}_{r,i} \right) \geq 1 \right) \right). \tag{4}$$

The length of a gap is less than s if the sum exceeds 1 in all windows of $s + 1$ columns that contain columns of the gap. An example illustration of the steps involved in the calculation of Equation 4 is given in Figure 4.

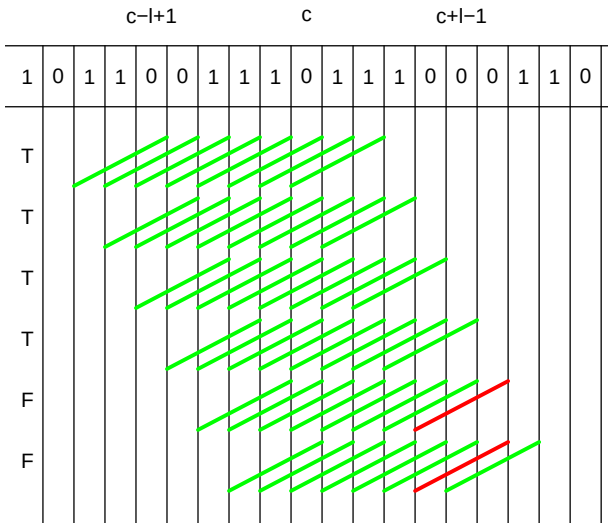

Figure 4: Visualisation of Equation 4 with parameters $l = 6$ and $s = 2$. An example sequence of $\mathbf{R}_r^{gd'}$ is given at the top and the positions of c and the bounds of the left and right most l long regions are marked above it. Each row of $s + 1$ long green and red lines represents the sums of the conjunction. Green lines represent cases where the sum ≥ 1 and the false cases are shown in red. The value of the conjunction is given on the left and the outer disjunction is true if any of these values is true.

Gap distance method can be implemented efficiently (the algorithm is given in Section 4.2) and it gives unambiguous results in both direction along the sequence. The distance parameter d, however, is extremely sensitive to uneven distribution of heterozygous SNPs making it quite hard to increase the parameter value. We improved the method by replacing the distance parameter with a sliding window that calculates the rate of heterozygosity around the SNPs.

Sliding window can be used to measure the level of local homozygosity at a particular subsequence. Sliding window is less sensitive to intermediate distances of genotyping errors in comparison to gap distance method. Let w be the length of the window and g the maximal heterozygosity rate and we can formulate matrix $\mathbf{R}^{sw'}$ as

$$\mathbf{R}_{r,c}^{sw'} = \left(\frac{\sum_{i=max(c-w/2,1)}^{min(c+w/2-1,k)} \mathbf{R}_{r,i}}{w} \geq 1 - g \right) \bigvee \mathbf{R}_{r,c}, \qquad w \in [1,k], 0 \leq g \leq 1. \tag{5}$$

Equation 5 gives us regions with lengths anything between 1 and k but too short regions can be filtered using the Equation 2:

$$\mathbf{R}_{r,c}^{sw} = \bigvee_{v=c-l+1}^{c} \left(\left(\sum_{i=max(v,1)}^{min(v+l,k)} \mathbf{R}_{r,i}^{sw'} \right) = l \right). \tag{6}$$

The last expression extends each region with the surrounding homozygous SNPs.

4.2 Detection Algorithms

Selection of the proper searching algorithm depends on the definition of the homozygous region. Some possible definitions were discussed in Section 4.1. The following three different methods have been implemented in CohortComparator: gap distance, gap ratio, and sliding window. All algorithms are quite similar in their structure and thus the same base algorithm can be adapted to all of them.

```
1  findHomozygousRegions(snps, l, s, d)
2  regions ← ∅
3  pS ← −1 : Start position of the region
4  pE ← −1 : End position of the region
5  gc ← 0 : Gap length counter
6  rc ← 0 : Region length since the last gap
7  for (col in snps) {
8      if (snps_col = NA) next
9      if (snps_col ≠ ab) {
10         if (pS < 0) pS ← col
11         rc ← rc + 1
12         gc ← 0
13         if (SNP_CODE) pE ← col
14     } else {
15         if (pS > 0) {
16             gc ← gc + 1
17             if ((gc > s) ∨ REGION_CODE) {
18                 if ((pE > 0) ∧ (pE − pS + 1 ≥ l))
19                     regions ← regions ∪ (pS, pE)
20                 pS ← −1
21             }
22         }
23         rc ← 0
24     }
25 }
26 if ((pS > 0) ∧ (pE > 0) ∧ (pE − pS + 1 ≥ l))
27     regions ← regions ∪ (pS, pE)
28 return(regions)
```

Figure 5: Template code for the recognition of homozygous regions. Blue parts of the code represent an example implementation of gap distance method. The implementations of the statements SNP_CODE and REGION_CODE vary in different region recognition algorithms. Gap distance method defines these statements as: SNP_CODE =($rc \geq d$) and REGION_CODE =($rc < d$).

Template of the generic region recognition algorithm is given in Figure 5. The region recognition function `findHomozygousRegions()` takes three parameters: a sequence of SNP genotypes (*snps*) of one sample, the minimum length of the region (*l*), and the maximum length of a continuous gap (*s*). A set of region start and end SNP pairs is produced for the given sequence. Method specific implementations of the algorithm vary in SNP_CODE and REGION_CODE statements and they have some additional bookkeeping variables in order to meet the requirements of these two Boolean statements. SNP_CODE is a Boolean statement that is true if and only if the homozygous SNP $snps_{col}$ can be accepted into a region. REGION_CODE is another Boolean statement, which is true if and only if the gap tolerance limits are exceeded by the heterozygous SNP $snps_{col}$.

Gap distance method is based on Equations 3 and 4. The algorithm keeps count of how many homozygous SNPs there has been since the last heterozygous SNP (*rc*) and SNP_CODE is a simple comparison whether $rc \geq d$. REGION_CODE can be given as a complement of SNP_CODE (rc $< d$).

Gap ratio method accepts heterozygous SNPs as long as the rate of heterozygosity is kept below the threshold *g*, where the parameter *g* is defined as a ratio of heterozygous SNPs since the opening of the region (pS in Figure 5) and the total amount of SNPs within that range is counted in *gt*. The parameter *gt* is incremented together with *gc* but reset to 0 along with pS. The generic template algorithm can be adjusted to the gap ratio method by replacing SNP_CODE with a check that $gt/(col - pS + 1) \leq g$. The same expression can be used in REGION_CODE except that "≤" operator has to be replaced with ">" operator.

The advantage of gap ratio method is that it is more flexible than the gap distance method in terms of variation in distances between the gaps and it is easier to determine an acceptable rate of heterozygosity than a distance limit. However, the gap ratio method has some major weaknesses. First, long homozygous regions are joined together although they have a significant amount of heterozygosity in between suggesting several separate regions. Secondly, long regions have a tendency of getting a tail of miscellaneous short repeats of homozygous SNPs. Thirdly, regions are interrupted by gaps in the beginning of the region. These drawbacks can be tackled by a sliding window that is used to determine the local gap density surrounding each SNP.

Sliding window method is the most complex of three homozygosity search methods used in CohortComparator. The implementation follows the concept of Equation 5 and the method has the advantages of the direction insensitivity and robustness against uneven distance between the genotyping errors. The generic template algorithm can be used for the sliding window by replacing SNP_CODE with a constant TRUE corresponding to the

disjunction of $\mathbf{R}_{r,c}$ in Equation 5. REGION_CODE is responsible for the left side of the disjunction where the ratio of heterozygous SNPs is calculated within the window and tested against the g.

locus:	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20
Sample 1:	aa	aa	aa	aa	aa	aa	aa	aa	aa	aa	aa	aa	ab	aa	ab	aa	ab	aa	aa	aa
gap distance	—	—	—	—	—	—	—	—	—	—	—	—	—	—				—	—	—
gap ratio L	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
gap ratio R	—	—	—	—	—	—	—	—	—	—	—	—						—	—	—
sliding window	—	—	—	—	—	—	—	—	—	—	—	—						—	—	—
Sample 2:	aa	aa	ab	aa	aa	ab	aa	aa	aa	ab	aa	aa	aa	aa	aa	aa	aa	aa	aa	aa
gap distance							—	—	—	—	—	—	—	—	—	—	—	—	—	—
gap ratio L							—	—	—		—	—	—	—	—	—	—	—	—	—
gap ratio R	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
sliding window							—	—	—	—	—	—	—	—	—	—	—	—	—	—

Table 2: A comparison of regions detected by different gap models. The parameters of the models are: l=3, d=3, g=0.2, and w=10. Homozygous regions are labelled with "—" and heterozygous regions are shown as blank.

Some of the differences between the methods of detecting homozygous regions are illustrated in Table 2. Gap ratio method is listed twice as its behaviour is direction dependent. Gap ratio versions L and R refer to iteration from left to right and vice versa, respectively. It can be seen that the gap ratio method is very sensitive to the gaps in the beginning of the homozygous region (Sample 2: locus 10) but tolerates many gaps after long homozygous regions (L: Sample 1, R: Sample 2). The context specific sensitivity of the sliding window is demonstrated in the recognition of the gap dense region at locus 13 of Sample 1, which is ignored by gap distance method. The locus 10 of Sample 2 is considered homozygous by both methods as it is surrounded by a continuum of homozygous loci.

It is worth of notice that gap distance, gap ratio, and sliding window all ignore the distances between the SNPs and assume a uniform distribution over each chromosome. The actual distribution of SNPs in contemporary microarrays is quite far from being uniform and the behaviour of the algorithm may vary in different parts of the chromosome [Nic06]. Reason for the variation is that all distances are measured in numbers of SNPs and thus the meaning of parameter values depends on the local SNP density. The compensation of non-uniform distributions has been implemented by providing an option to use base pair distances in parameters l and w. The implementation of this option required an additional input array defining the SNP loci and the distance calculations between

pS, pE were conducted using the array values of corresponding indices. The bounds of the sliding window are calculated by searching the outer most SNPs within the range of $[col - w/s, col + w/2 + 1]$.

Hidden Markov models are widely used statistical models for detecting sequence types in bioinformatics. The models consist of states, transition probabilities between the states and emission probabilities for having a certain value in each state. The distinction between homozygous and heterozygous regions could be modelled in various ways. In the simplest case there would be two states $X_{homozygous}$ and $X_{heterozygous}$, transitions $A_1 = X_{homozygous} \rightarrow X_{homozygous}$, $A_2 = X_{homozygous} \rightarrow X_{heterozygous}$, $A_3 = X_{heterozygous} \rightarrow X_{homozygous}$, $A_4 = X_{heterozygous} \rightarrow X_{heterozygous}$, and the emission probabilities of getting a homozygous or heterozygous SNP for both states. Even the simplest HMM contains two transition probabilities and four emission probabilities that should be estimated somehow. The estimation of these parameters could be done using Baum-Welch algorithm [Bau70] in case there are samples with known homozygous and heterozygous regions but the information is not available as the labelling of the regions depends on the definition of homozygous region itself.

The HMM models tend to produce ambiguous bounding for the homozygous regions depending on the iteration direction along the SNP sequence. The ambiguities are caused by Viterbi algorithm [Vit67] that is used to find the most probable sequence of states for the observed sequence of emitted values (SNP genotypes). The algorithm keeps list of probabilities of possible state sequences during the iteration of the genotypes and selects the most likely transition based on the transition probability and the probability of emitting the observed value in that state. Emission probabilities of homozygous SNPs in $X_{heterozygous}$ and heterozygous SNPs in $X_{homozygous}$ represent the rate of genotyping errors [Ber06]. Transitions between homozygous and heterozygous regions (A_2, A_3) are less probable than those between the states (A_1, A_2). The differences between the results of the different iteration directions depends on how many false emissions can be accepted before it becomes more likely that the emissions are true but the underlying state has changed.

The HMMs have been used for the detection of LOHs in comparison of “normal” and tumour samples [Lin04] and plain tumour samples [Ber06]. Beroukhim et al. [Ber06] compared their HMM against a simpler model that they called NumHom. NumHom is essentially what is captured by $\mathbf{R}^l$ and it was shown to be less specific than the HMM. The lower specificity was caused by the stretches of homozygous SNPs in linkage disequilibrium; the stretches that were are trying to capture in this study.

4.3 Scoring

The genome wide analysis of a genotype data set may give more than thousand possibly interesting regions depending on the parameters and number of samples. The complete analysis of all these regions for all associated samples would take far too much time and resources. Clearly, a scoring scheme is required for the set of possibly interesting regions.

The scores can be calculated only if there is a metric that tells, which are the most promising candidates for the wet lab analysis. Qualities of the regions can be described in terms of the length, relative frequency of its features between the controls and the patient samples [Woo04], and its biological content. In this section we concentrate on the first two properties also used in CohortComparator. The last and perhaps the most important property is discussed in Section 5, where we look at the annotations.

A possibly interesting region consists of samples with homozygous alleles overlapping the region loci. The region consists of alleles that are not homozygous in reference samples. The type of the interesting region depends on the number of alleles that it covers. An interesting region may represent a feature of general homozygosity if there are not too many overlapping homozygous regions in references. Interesting regions may represent allele specific features if they have an allele that is not homozygous in references but the references are homozygous for some other alleles. An interesting region may cover the whole length of a homozygous region in one sample but another sample may share a fraction as well. The scoring of the region should be based on the length of the features and thus we are not calculating the actual length of the region itself but the total length of sample features overlapping it. CohortComparator implements two different methods for this: `total` includes the length of the whole feature even if a part of it overlaps with the region; `fraction` takes only the intersection of the feature and the region into account. The difference between these two methods is visualised in Figure 6, in which two interesting regions `A1` and `A2`. The features are required from two or more samples in this example and they must not be present in the reference. `A1` consists of blue alleles present in homozygous regions `r3`, `r4`, and `r8`. `A2` represents the homozygosity of homozygous regions `r5`, `r7`, and `r9`. Parts of the homozygous regions that are used in the scoring methods, `total` and `fraction`, have been marked with yellow and green lines, respectively.

Both scoring schemes have been tested in genome wide analysis with the same parameters and the `fraction` method agrees better with the manual scoring given by a biologist. Long homozygous regions dominate the results of the `total` method even if the interesting region is very short and shared by few individuals.

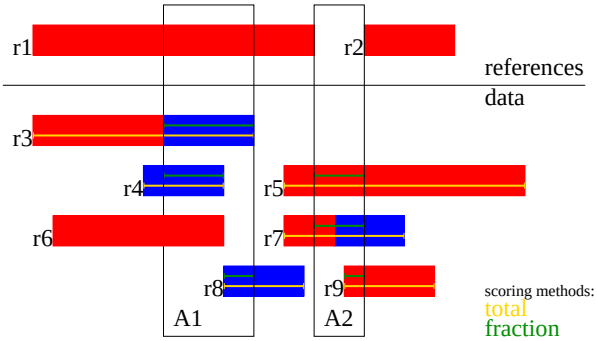

Figure 6: Example pattern of homozygous regions.

Score values are most useful as sorting criteria. The interpretation of the meaningfulness of the value itself is quite hard since the range of possible scores is $[0, k]$. Getting the score of k would mean that all SNPs would differ in all samples between the data sets. The analysis of the score distributions revealed an interesting fact that a function of $score(regionNumber) = a/regionNumber$ fitted extremely well to observed values. Parameter *regionNumber* is the index of the interesting region within a sorted list of scores, whereas a refers to a scaling factor that is calculated by minimizing the function $\sum |score_i - (a/i)|$. Visual comparison of the scores and the function is given in Figure 7. The estimation function $a/regionNumber$ of the scores is monotonically decreasing but non-linear within [1, $max(x)$]. One possible way of determining meaningful scores is to calculate the point of the curve after which the decrease is too slow that is the second derivative of the curve is greater than a given limit. Another method for the estimation of saturation point of the scores is to take an advantage of the symmetry of the $1/x$ function. The score density (y-axis) is higher than the region sample density (x-axis), when the scores are below the intersection of the symmetry axis if both axes are scaled to the same range, say [0, 1]. The meaningfulness of the score values is greater in regions exceeding the limit whereas other consecutive (based on their scores, not physical locations) regions have little or none difference in their scores. The scaling factor for the axis is estimated as $max(score)/count(regions)$.

4.4 Compound heterozygosity

Regions of *compound heterozygosity* contain two different mutant alleles at the same locus. Individuals with compound heterozygote regions may express recessive phenotypes if both alleles have the same or similarly behaving mutation. Compound heterozygotes

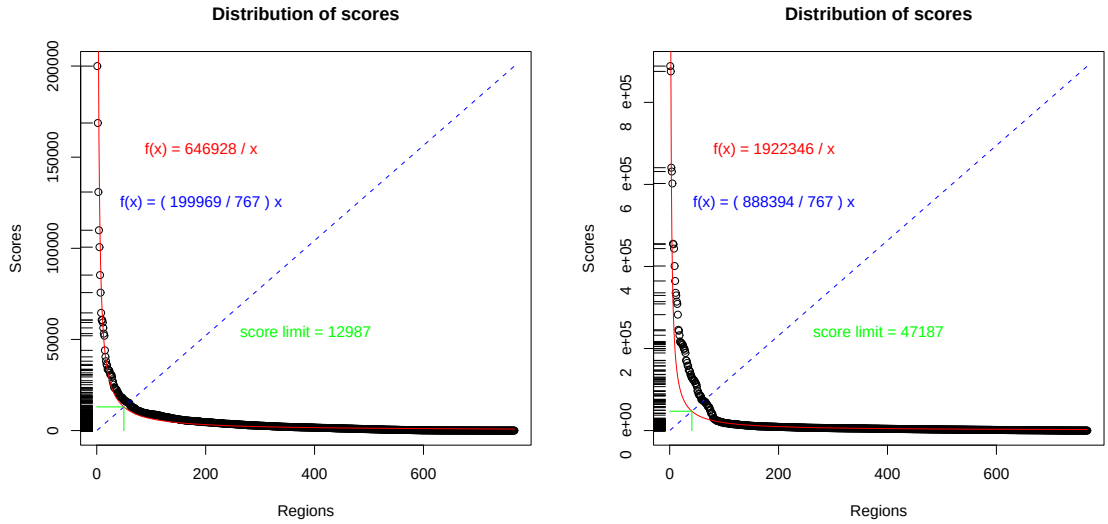

Figure 7: Score distributions of `fraction` and `total` methods. Vertical bar on the left contains a black line if there are interesting regions with the corresponding score giving a broad overview of score densities. Exact scores are plotted in descending order so that x-axis represents the index of each region within the list. The best fit of the red a/x curve and its intersection with $(max(score) \times x)/count(regions)$ is used to estimate the saturation point of scores. Left diagram is created using `fraction` scores and the right one is based on `total` scores. Scores of the figure are from the genome wide analysis with parameters: $l.limit.d = 55, l.limit.r = 25$, sliding window $(w = 20, d = 0.05)$, $f.limit.d = 0.03$, $f.limit.r = 0$.

can be affected with recessive disorders although their parents are not related [MYE98].

Regions of potential compound heterozygosity for alleles in **R** can be defined as:

$$\mathbf{C}_{r,c} = \begin{cases} 1, & \bigvee_{s=1}^{n} \Big((\mathbf{S}_{r,c} = \underline{ab} \vee \mathbf{S}_{r,c} = \mathbf{S}_{s,c} \vee \mathbf{S}_{s,c} = \underline{ab}) \wedge (\mathbf{R}_{s,c} = 1) \Big) \\ 0, & \bigvee_{s=1}^{n} \Big((\mathbf{S}_{r,c} = \underline{aa} \wedge \mathbf{S}_{s,c} = \underline{bb}) \vee (\mathbf{S}_{r,c} = \underline{bb} \wedge \mathbf{S}_{s,c} = \underline{aa}) \vee (\mathbf{R}_{s,c} = 0) \Big), \end{cases} \tag{7}$$

where $r \in [1, n]-s$. The matrix **C** contains 1 for each SNP that may be a combination of an allele in any overlapping homozygous region in **R**. The definition as such is obviously too weak since a huge amount of short sequences is generated simply by chance. A simple and consistent method to increase the reliability of the regions is to set the minimum length (l) for the region of interest as was done with the homozygous regions in Equation 2. The replacement of the homozygosity matrix **R** with **C** gives us a more informative matrix:

$$\mathbf{C}^{F}_{r,c} = \bigvee_{v=c-l+1}^{c} \left(\left(\sum_{i=max(v,1)}^{min(v+l,k)} \mathbf{C}_{r,i} \right) = l \right), \qquad l \in [1, k]. \tag{8}$$

The construction of the matrix $\mathbf{C}^F$ in Equation 8 does not take SNPs into account if they are not part of a homozygous region and it assumes that each homozygous region is a

continuous haplotype. The content of a homozygous region is, however, a sequence of haplotypes that might have a different origin. Pieces of different pre-existing haplotypes may be found from a single homozygous region and thus another individual may be compound heterozygote for parts of the region sharing some of the original haplotypes. A long homozygous region compared against another sequence, where the length of the sequence is much less than l, gives us those partial compounds that are not too close to the ends of the homozygous regions. The detection of compound heterozygotes at the ends would require a check for $l - 1$ SNPs before and after the region. Problem with these SNPs is that we do not know the chromosome specific sequences as the SNPs may be heterozygous. The use of heterozygous SNPs would require information about actual DNA strands responsible for the SNP alleles in order to create sequences for the compound search.

An estimation of the DNA strands can be done by calculating an estimate of the haplotypes of the sample population (see fastPHASE in Section 6.2). Compound heterozygotes can be determined to any chromosomal region if the haplotypes of all samples are known. Compound heterozygotes can be found by checking whether a sample contains the mutant alleles at the given locus. An important difference to the method of detecting potential sites of compound heterozygosity described above is that both strands can be compared in full length. Possibly interesting alleles can be found although they are not homozygous in any of the samples but if they are seen in compound heterozygotes together with an allele that has been homozygous but such allele combinations are absent in reference set. Chances are that both allele contain a similarly behaving mutation and thus express a recessive phenotype (=developed a cancer). The use of haplotype data in the search of compound homozygotes is out of the scope of this study and will be incorporated to forthcoming releases of CohortComparator.

The algorithm used for the recognition of the potential compound heterozygotes is almost identical to that used in the recognition of homozygous regions (see Figure 5). The only exception is that $snps_{col}$ is not compared against $\underline{ab}$ (line 8) but against the given homozygous region ($comp$) that is supposed to match the other component of the compound. The comparison expression is $(comp_{col} = NA) \vee (comp_{col} = snps_{col}) \vee (\underline{ab} = snps_{col}) \vee (\underline{ab} = comp_{col})$. The only purpose of the statement $\underline{ab} = comp_{col}$ is to allow both alleles where a gap was found in the original homozygous region. The homozygous regions can be represented as haplotypes of two identical strands of homozygous alleles. The haplotype representation of the homozygous region would lead to a theoretically more sophisticated formulation in a sense that it is a haplotype that is searched for but the results would remain the same. The haplotype representation of a homozygous region would contain

missing values (NA) in places of heterozygous SNPs and the comparison would be given as $(comp_{col} = NA) \vee (\underline{ab} = snps_{col}) \vee ((comp_{col} = \underline{a}) \wedge (snps_{col} = \underline{aa})) \vee ((comp_{col} = \underline{b}) \wedge (snps_{col} = \underline{bb}))$. In this context, the use of haplotype representation is not computationally feasible since the allele combinations are readily available in matrix **S**.

Parameters of the algorithm are chosen so that each violation of the given expression terminates the region. A region cannot be considered as a possible compound heterozygote only if there is a homozygous SNP with a complementary allele. The probability of having l SNPs without such alleles is already so high that false positives can be expected but the addition of gaps into this model would increase their number even more. The gap distance method with $d = l$ is used since it the fastest algorithm. An additional optimisation is done concerning the last $l - 1$ SNPs. There is no need to check last SNPs if $pE < 0$ and the loop can be terminated. The same trick could be used in the search of homozygous regions but the code has been coupled with the detection of deletions that can be found from the tail of the sequence.

Equation 7 defines all regions that are possible compounds of any homozygous region in **R** but the number of regions can be significantly decreased if the matrix is constructed so that it includes only those regions of the cancer patients that overlap with an interesting region they belong to. The final results are clipped to the interesting areas as the result of such limited search consists of compound heterozygotes that are closely around the interesting regions but the exact bounds depends on the length of those areas of the homozygous regions that are outside of the interesting region. All interesting areas with their homozygous regions are analysed in the order of their loci, which means that the same homozygous region would be analysed several times if it belongs to several interesting areas. Comparing this region against the others would not be feasible if that has been done already and thus a list of end loci of samples is kept. The end position of the homozygous region is compared against this list each time a homozygous region is selected from an interesting region. The actual search takes place only if the value associated to the sample is less than the new value and the sample value gets updated.

The interesting regions that contain no compound heterozygotes are of the special interest since that is a sign that the alleles of the patient regions are especially rare within the population. CohortComparator reports all such regions in its output and the regions are emphasized in its graphs (see Appendix 1). The current version of CohortComparator does not report compound free regions with reference compounds to one or more homozygous allele, although there would be other alleles that do not have compound pairs in references.

5 Annotations

The list of interesting homozygous regions consists of sequence combinations that are rare in references. However, it contains no information about biological effects caused by this variation. Further, the scoring scheme cannot be used to distinguish cancer related regions from other homozygous regions. The biological significance of a genomic region depends on its cellular function in terms of transcripts, protein binding affinities and folding (3-dimensional structure) properties. A mutation in a critical part of the DNA may lead to significant changes in individual's phenotype; a gene may get inactivated, for instance. The chance of developing a cancer increases if a tumour suppressor gene gets inactivated in both chromosomes. It is not know, which genes are tumour suppressors, as we are trying to find new ones. Locations of the most genes are, however, known. Information about the genes can be assigned to interesting homozygous regions and it can be assumed that some of the genes may be involved in the development of cancer. On the other hand, a region is less likely to be worth a wet lab investigation if it does not have a gene association. A high number of genes found to overlap with a region does not necessarily make the region more interesting since an interesting mutation may affect just one gene. The attention can be focused on regions with cancer associated genes but one must keep in mind that almost all genes are somehow associated to cancer and the interpretation of the relationship may be hard to do in an automate manner.

The biological content of the interesting regions is retrieved from Ensembl genome database [Bir06], which contains DNA sequence and gene information of various animals. The database version used in this study covers 21571 known human genes, 2142 novel genes waiting for confirmations and a variety of RNA transcripts. The database content is maintained in regular basis as the biological knowledge alters and the content is publicly available for direct database queries and database exports that can be copied to the local replicates. Ensembl database describes each transcript with a variety of properties such as:

1. a *common name* for the gene,
2. the *chromosome location*,
3. a *list of exons* and their locations,
4. *type of the gene* discriminates between the protein coding genes and various types of RNAs,
5. a *list of transcripts* and splice variants encoded by the gene,
6. *status of the confidence* that tells whether the gene is well known or predicted by bioinformatical software for example,

7. a short *description* of its biological function,
8. *references to microarrays* targeting the gene, and
9. *database identifiers* for other databases.

The list possibly interesting chromosomal regions is compared against Ensembl database and a list of overlapping genes is associated to each region. An example list or genes associated to interesting regions is shown in Table 3. The list can be used for several important purposes. First, the lengths of the genes and the exon counts can be used to estimate how much work is needed to sequence the possible mutations. Second, the microarray probe identifiers help in comparing the gene list against the gene expression studies if such information is available. Third, the list of possibly affected genes can be compared against other databases describing other interesting aspects of genes.

```
.
.
- 4:  88155377-88277512   (5816)
  4:  88075187-88281214   AFF1 (KNOWN): protein_coding, AF4/FMR2 family member 1 (Protein AF-4) (Proto-oncogene AF4) (Protein FEL). [Source:Uniprot/SWISSPROT;Acc:P51825], 24 exon(s), probes: [U133_X3P:214448_3p_x_at:1109:501;] IDs: [accession=ENSG00000172493]
- 3:  161313644-161435642 (5809)
  3:  161426323-161428693 Q8N5S4_HUMAN (KNOWN): protein_coding, 1 exon(s), probes: [U133_X3P:201555_3p_at:1140:647;] IDs: [accession=ENSG00000180044]
  3:  161427938-161650320 TRIM59 (KNOWN): protein_coding, Intraflagellar transport 80 homolog (WD repeat protein 56). [Source:Uniprot/SWISSPROT;Acc:Q9P2H3], 28 exon(s), probes: [U133_X3P:Hs2.105633.2.S1_3p_a:649:1141;] IDs: [accession=ENSG00000068885]
- 6:  49771426-49852590   (5797)
  6:  49768032-49789258   CRISP2 (KNOWN): protein_coding, Cysteine-rich secretory protein 2 precursor (CRISP-2) (Testis-specific protein TPX-1). [Source:Uniprot/SWISSPROT;Acc:P16562], 14 exon(s), probes: [HG_U95Av2:34480_at:61:247;] IDs: [accession=ENSG00000124490]
  6:  49803056-49820040   CRISP3 (KNOWN): protein_coding, Cysteine-rich secretory protein 3 precursor (CRISP-3) (SGP28 protein). [Source:Uniprot/SWISSPROT;Acc:P54108], 11 exon(s), probes: [HG_U95E:70801_at:370:17;] IDs: [accession=ENSG00000096006]
- 15: 46920822-47004014   (5768)
  15: 46903227-47042933   SHC4 (KNOWN): protein_coding, rai-like protein [Source:RefSeq_peptide;Acc:NP_976224], 12 exon(s), probes: [HG_U95E:76135_at:608:285;] IDs: [accession=ENSG00000185634]
  15: 46957582-46959672   NM_014335.2 (KNOWN): pseudogene, CREBBP/EP300 inhibitor 1 (CRI1), mRNA [Source:RefSeq_dna;Acc:NM_014335], 2 exon(s), probes: [HG_U95B:52746_at:434:143;] IDs: [accession=ENSG00000178558]
- 2:  80998358-81118937   (5742)
- 8:  78328040-78448532   (5738)
- 1:  105812910-105933049 (5721)
- 8:  14623177-14743200   (5715)
  8:  13991744-15140219   SGCZ (KNOWN): protein_coding, Zeta-sarcoglycan (Zeta-SG) (ZSG1). [Source:Uniprot/SWISSPROT;Acc:Q96LD1], 10 exon(s), probes: [U133_X3P:Hs.123349.0.A1_3p_at:487:1009;] IDs: [accession=ENSG00000185053]
- 3:  163227593-163287964 (5715)
- 13: 48209948-48329829   (5709)
  13: 48237029-48237137   5S_rRNA (NOVEL): rRNA, 5S ribosomal RNA [Source:RFAM;Acc:RF00001], 1 exon(s), probes: [HG_U95Av2:378_s_at:392:491; HC_G110:378_s_at:246:263;] IDs: [accession=ENSG00000199376]
.
.
```

Table 3: Annotation list of interesting regions Interesting regions are listed in the descending order based on their scores. Each region starts with a hyphen followed by its genomic location given in a form of chromosome number and the range of base pairs. The location information is followed by the score, which is given in parenthesis. The biological content of the region is given in below each region so that each transcript is listed in its own line starting with the associated chromosome location. The name of the transcript is followed by its status and biological type. A short description of the gene is given before an exon count and the lists of associated microarray probes and database identifiers.

Gene Ontology (GO) [Ash00] is a database that consists of terms representing biological concepts such as 'apoptosis' and 'response to virus'. The terms have been organised into a directed acyclic network that describes their relationships so that the terms are bind to their more generic parents. For example 'apoptosis' is a child of 'programmed cell death', which in turn is a child of 'cell death' and so on. RegionAnnotator provides a query tool that can be used to select interesting regions based on a GO term. The program takes RegionAnnotator output files as its input and compares all GO.

SNPs3D database [YMM06] can be queried for disease associated genes. Disease Candidate Gene module of the database provides an interface to a text mining tool. The tool can be used to extract gene names based on keywords that are used in the abstracts of articles available in Medline. The set of keywords is constructed by selecting abstracts with the name of the disease. The words of these disease related abstracts are compared against all abstracts to filter out common words that are not related to disease. The 40 terms that are the most restricted to the scope of the disease are used as keywords. For example the set of colorectal cancer keywords suggested by SNPs3D site consists of such words as:*"colorectal cancer", "hnpcc", "crc", "colorectal", "nonpolyposis colorectal", "hereditary nonpolyposis", "nonpolyposis", "cancer hnpcc", "msh2", "colorectal neoplasms", "colorectal neoplasms, hereditary nonpolyposis", "colorectal cancers", "hmsh2", "mlh1", "hmlh1", "msi", "sporadic colorectal", "mmr genes", "cancer crc", "colorectal cancer patients", "human colorectal", "non-polyposis colorectal", "hereditary non-polyposis", "non-polyposis", "amsterdam", "colorectal cancer cells", "mmr", "msh2 protein, mammalian", "mlh1 protein, mammalian", "microsatellite instability", "colorectal cancer cell lines", "primary colorectal", "amsterdam criteria", "hereditary nonpolyposis colorectal cancer", "mismatch repair", "hereditary nonpolyposis colorectal cancer hnpcc", "dna mismatch", "hnpcc patients", "repair mmr", and "instability msi"*. SNPs3D retrieves the gene names from NCBI Gene database and compares them against the disease articles. The genes that are mentioned in articles are returned with an association score indicting how well the keywords and the gene name matched the abstracts. The list of genes received from SNPs3D database can be compared against the region annotations of RegionAnnotator.

Interesting lists of genes can be found not only from the databases but also from some articles. For example, an extensive study of colorectal and breast cancers was explained in [Sjö06]. The supplement lists of mutated genes in colorectal cancers provided a source of potential genes that may have affected in the development of the cancer.

6 Materials and Methods

The cause of the CRC is likely to vary between the patients and we expect to see only few individuals sharing a common recessive mutation [Aal]. The practical consequence is that we must be able to detect 2–5 positive hits out of some tens of obtained candidates making the signal far too weak for standard statistical tests between the patient and reference samples. The problem can be seen when the allele distributions are compared between the sample groups and false discovery rate correction [PC06] filters out all possibly interesting loci (see Section 7). Long continuous homozygous regions have been used to identify regions that are likely to be autozygous.

6.1 Origin of the data

Samples of the colorectal cancer patients ($\mathbf{S}^D$) were all from Finland. Total of 50 unrelated patients have been selected out of 1044 colorectal cases (the original collection of colorectal cancer patients is described in [Aal98, Sal00]). The cause of the cancer in these 50 patients was not explained by known reasons and they had at least one sibling with a colorectal cancer, which suggests a recessively inherited mutation. The parents of the patients were both healthy. The samples with a known cause (microsatellite instability caused by a mutation in mismatch repair genes such as *MLH1*, *MSH2*, *MYH* and *MYH11*) were excluded as we are trying to detect new tumour suppressor genes.

DNA was extracted from the frozen tissue samples of the selected 50 patients were sent out to a lab [4] in the US for the hybridisation. The microarray experiments were carried over using Affymetrix GeneChip®Human Mapping 100K Set and the standard protocol [Aff04]. The chips were scanned and the raw data was sent back to us in Affymetrix `.CEL`, `.CPH`, and `.EXP` files.

Samples with more than $0.05k$ `NoCall` SNPs were considered unsuccessful since they would have too many false positives in homozygosity analysis. Unsuccessful cases were left out of studies leaving 42 sets of data for the analysis. The sample material was not collected from tumours but the surrounding healthy tissue and thus we are not analysing colorectal cancer genomes but genetic background of those subjected to it.

Reference samples ($\mathbf{S}^R$) were of the miscellaneous origin but they all were analysed using only one (XbaI) of the two microarrays. The consequences of the lower SNP resolution,

[4] The hybridisation was carried over by a company called Expression Analysis, Inc. (Durham, NC, US). Www pages for more information about products and services: http://www.expressionanalysis.com/.

such as false positives in homozygous regions and differences in SNP densities, are further discussed in Section 7. The sample donors were not close relatives of each other. Most reference samples (41) are from The Finnish Red Cross (FRC) and the rest 10 samples were from the previous studies of Department of Medical Genetics. The samples from FRC were part of their leukaemia studies and they have been takes from healthy siblings of leukaemia patients.

The genotyping was performed for all chromosomes but only autosome data has been used in our study. Sex chromosomes have been left out from the studies because male samples have a single copy of both chromosomes X and Y and females have a diploid composition of chromosomes X. Mutations in chromosome X would affect in phenotypes of males even if they are of recessive type and such hemizygous effects can be seen right from the pedigrees without advanced analysis. Chromosome Y could be analysed from male samples only, which would reduce the amount of available data. The chromosome has been excluded because of the small number of samples and their constant hemizygosity.

AffyExportParser (described in Section 6.2.1) was used to convert sample data into form of genotype matrices (Figure 10) used by statistical software. SNPs without a known physical locus (466 SNPs from XbaI chip and 385 SNPs from HindIII chip) and those found in sex chromosomes were dropped during the conversion. The total amount of SNPs left for the study was 57290 (XbaI) + 55700 (HindIII) = 112990 for the patients and 57290 for each reference sample.

6.2 Programs

The conversion from SNP-microarray measurements to biological knowledge involves a series of computational procedures implemented in various programs. In this section we describe the software we have been using in this study. The exact versions of databases and software distributions are listed in Table 4.

A schematic view of the data flow is given in Figure 8. The initial SNP data comes in a platform specific format and we start by converting it into a generic form of chromosome size matrices of two allele genotypes. The conversion was done in several steps using Affymetrix software for the extraction of genotypes and our own code (analysis steps) to merge all chips and individuals into matrices.

The data analysis starts by importing the hybridisation measurements of each sample into a local database. Affymetrix Data Transfer Tool (DTT) reads Affymetrix `.CEL`, `.CPH`, and `.EXP` files created during the array scanning and restores them into a pool of all

Program	version
Affymetrix Data Transfer Tool	1.1.0
Affymetrix GTYPE	4.0
CohortComparator	1.3
Ensembl genome database	40.36b
fastPHASE	1.1.4
JAVA™2 Platform Standard Edition	1.5.0
R	2.4.0

Table 4: Version numbers of databases and software used in this study

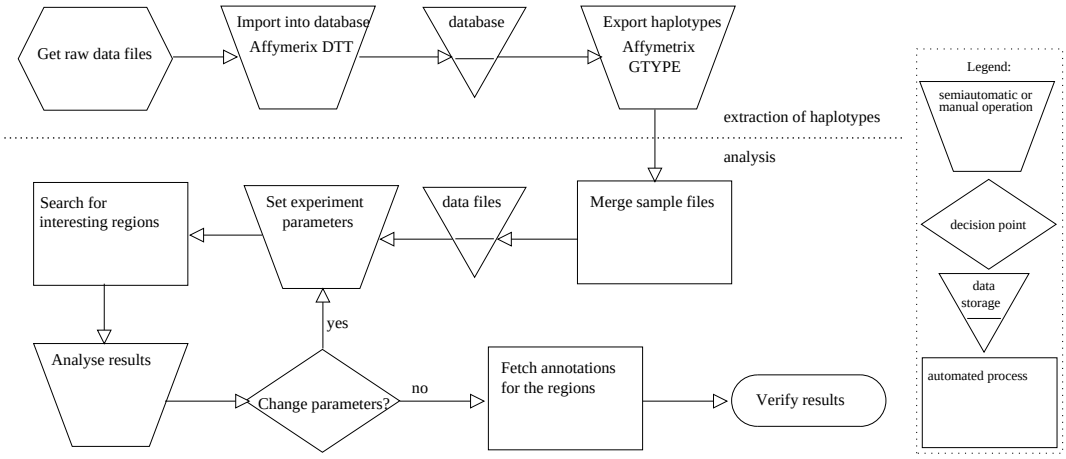

Figure 8: Steps of the data analysis.

experiments. The local database can be used along with Affymetrix software but the vendor specific storage is difficult to access for external applications. Genotypes can be extracted from the database by using Affymetrix GTYPE genotyping program, which takes care of probe annotations, background corrections and normalizations of the data.

Probe annotations are loaded automatically from the Affymetrix web site and they are used to identify feature sequences, to features into SNP alleles and associated locations in human genome. The background correction is done by comparing the signals of the perfect match probes to those of mismatch probes.

Figure 9 illustrates the format of the Affymetrix GTYPE export being a tab delimited text file consisting of rows of SNPs and their attributes. The first two lines are headers describing the file title and attribute names.

```
Dynamic Model Mapping Analysis
     SNP ID        Chromosome Physical Position dbSNP RS ID      TSC ID     EA05021...Call EA05021...Confidence
1    SNP_A-1746877 0                                                        BB             0.000488
2    SNP_A-1709473 0                                            TSC64405    AA             0.002930
3    SNP_A-1686560 0                                            TSC648024   AB             0.000488
4    SNP_A-1731395 0                                            TSC1587919  BB             0.126953
5    SNP_A-1685046 0                                            TSC393273   AB             0.005859
6    SNP_A-1707717 0                                            TSC618160   AB             0.000488
7    SNP_A-1692726 0                                            TSC1216283  AA             0.000488
.
.
.
1169 SNP_A-1658184 1          58519238               rs3861824  TSC1332643 NoCall         0.669922
1169 SNP_A-1658184 1          58519238               rs3861824  TSC1332643 NoCall         0.669922
1170 SNP_A-1754126 1          58561256               rs7367176              BB             0.000488
1171 SNP_A-1663466 1          58570387               rs10489813             BB             0.000488
```

Figure 9: Affymetrix GTYPE export file example.

6.2.1 AffyExportParser

AffyExportParser is a Java program that takes its input in the form of Affymetrix GTYPE export files (Figure 9) and produces a set of tab separated text files that can be used in R [R D06] and other statistical software. AffyExportParser is especially suitable for platforms with two chips because it is capable of merging SNP measurements of both chips into a single vector but it can be used as a standard tool for the input file construction even on one chip platforms.

AffyExportParser expects that the content of the input files is sorted by chromosome number and physical location of SNP. The sorting can be done in Affymetrix GTYPE before the data export and thus it has not been integrated into AffyExportParser. SNPs are all in the same order in unsorted export files and thus it is possible to sort them in $O(n \times k)$ time using a constant list of indices of input rows in the output order. The memory requirement of such algorithm would be $O(k)$ if we first read the whole sample data into main memory and then print them in the order of the index list. The time efficiency can be compromised for the sake of better memory efficiency depending on the amount of free memory and the size of k but that should not become an issue with contemporary computers and microarrays of size $k \lessapprox 10^6$.

The time and memory efficiency has been a priority during the implementation of AffyExportParser. Each chip data input file gets scanned just once and there are no more than two such files (in case of file merging between two chips) open at a time. The files are read line by line (=SNP by SNP) and merge algorithm is applied so that only one line is kept in memory for each file. The memory consumption of the algorithm is obviously $O(k)$, while the time consumption is $O(n \times k)$.

Each Affymetrix GTYPE export file contains all SNPs of the chip concerned but the analysis is usually conducted in a chromosome specific manner. AffyExportParser works

so that it keeps track of the chromosome that it is processing and the output file for the genotypes is changed according to it. New files are created during the iteration of the first sample and the next samples are appending into these files. A partial example of such output file can be seen in Figure 10. The first column contains the sample name and the following tab separated columns represent the corresponding haplotype combinations (A=aa, X=ab, B=bb, N=missing value). All SNPs are in order based on their physical locations, which are stored into separate files. The locus files are simple chromosome specific lists of base pair rows in the same order as the matrix columns. Locus lists are created during the iteration of the first sample but they are used by all other iterations to double check that the order of SNPs is identical.

```
C1001 A B B A A A B B B A X A A A A A B B X X X B B B B N B B ...
C100  A B B A A A B A B X X A A A X A B B A A B B A B B B B B ...
C135  A B B A A A B X B A A A A A B A B B X A X B B B B B B B ...
C182  A B B A A A B X B A A A A A A X B A X X X X B B B X X B ...
C205  A B B A A A B X B A A A A A X A B X X X X X B B B X B B ...
C206  X B B A A A B X B X X X X A X A B X A X B X B B B X N N B ...
C216  A B B A A A B B B A A A A A B A B B X A X B X N X A B B ...
⋮
```

Figure 10: An example of AffyExportParser output matrix.

The fact that we had only half as dense SNP set for the reference samples was causing problems in detecting the homozygous regions and fine tuning the parameters as described in Section 7. One way of reducing the number of false positives is to use haplotype data for the estimation of the missing SNPs. The linkage between neighbour SNPs is quite high and the values of the missing SNPs are likely to match the SNPs of those haplotypes of the population that are compatible with the SNPs around the missing values. The process of estimating the haplotypes differs depending on whether the samples are taken from relatives with a known pedigree or only distantly related individuals of the same population.

6.2.2 fastPHASE

A program called fastPHASE written in University of Washington can be used to estimate haplotypes from SNP data sets [SS06]. The major advantage of fastPHASE is in its scalability. The authors of fastPHASE have also published another haplotyping program called PHASE [SD03] but the amount of data we had (for example in chromosome 2 there were 10339 SNPs in 93 samples) was too much for it. The processing of all chromosomes took only some hours using fastPHASE.

Two data conversion programs were written in order to use fastPHASE: one that converts AffyExportParser output matrices to fastPHASE input files and another one for the reverse conversion of the haplotype outputs. The programs can be used for the construction of an additional set of data files (after the step “Merge sample files” in Figure 8) if the resolution of the reference and data sets differs. The original AffyExportParser outputs should be kept for the analysis of the deletions and effects of the missing value estimation.

6.2.3 CohortComparator

CohortComparator is a program that can be used for the actual comparison between the patient and reference samples. CohortComparator detects the homozygous regions (see Section 4.1) from each sample, separates alleles of the regions, compares $\mathbf{R}^D$ to $\mathbf{R}^R$ and calculates scores for the differences. CohortComparator produces a PostScript image (see Appendix 1) of each chromosome and a list of interesting chromosomal regions.

CohortComparator is technically a set of R scripts that are combined into an application. A special code library called AsserTools encapsulates almost all generic procedures that are used within the main program. The separately organized code library has been written in a recyclable manner so that it can be used together with other R programs. AsserTools provides its own framework for the unit testing. Unit tests are used to verify the procedures provided by the code library itself but the framework is readily useable for other R scripts as well.

A pseudocode of CohortComparator is given in Figure 11. The pseudocode does not explain the details nor naming conventions of the original source code but it follows the logical structure of the algorithm. The functions of the pseudocode are explained in the next paragraphs.

`ReadConfigurations()` reads the execution parameters such as the folder and file name options for the input and output files. The parameters of analysis, the method of score calculation, and the definition of the homozygous region are also read during the start. A complete list of the runtime configuration of CohortComparator can be found from Appendix 2.

```
1  ReadConfigurations()
2  results ← ∅
3  members ← ∅
4  for (chr in chromosome.set) {
5      loc ← LoadPhysicalSNPLocations(chr)
6      S^D ← LoadPatientData(chr)
7      S^R ← LoadReferenceData(chr)
8      R^D ← HomozygousRegions(S^D, loc)
9      R^R ← HomozygousRegions(S^R, loc)
10     homozygousRegions ← free(R^R) \ free(R^D)
11     t ← TestNormality(R^D)
12     col ← SetAlleleColors(R^R)
13     R^Da ← AlleleRegions(S^D, R^D, col)
14     R^Ra ← AlleleRegions(S^R, R^R, col)
15     alleleRegions ← free(R^Ra) \ free(R^Da)
16     chrResults ← homozygousRegions ∪ alleleRegions
17     chrMembers ← RegionMembers(chrResults, R^Da)
18     C^D ← CompoundRegions(chrResults, chrMembers, S^D, S^D)
19     C^R ← CompoundRegions(chrResults, chrMembers, S^D, S^R)
20     results ← results ∪ chrResults
21     members ← members ∪ chrMembers
22     : Create chromosome specific outputs
23     : graph (see Appendix 1) and
24     : a text file for region loci.
25     PrintChromosome(chr, t, col, loc, R^Ra, R^Da, C^R, C^D)
26 }
27 results ←MergeRegions(results)
28 results ←SortByScores(results)
29 PrintInterstingRegions(results)
```

Figure 11: Logical structure of CohortComparator.

`LoadPhysicalSNPLocations()` is a simple text parsing routine that converts the chromosome specific SNP locus lists of the reference and the patient data into lists (*loc*) that can be used together with **S** and **R** matrices. The locus list files are created by AffyExportParser and the order of loci follows the order of columns in sample matrices. The sample matrices are loaded using `LoadPatientData()` and `LoadReferenceData()`.

`HomozygousRegions()` is essentially a loop that calls `findHomozygousRegions()` function (template code is shown in Figure 5) for each sample (row) of the given matrix. The locus information is required only if the lengths (l, w) are given in base pairs. The recognition of homozygous regions and deletions are both done during the same allele scan in CohortComparator but the code related to deletions is omitted for simplicity.

The chromosome regions containing less than given amount of overlapping regions (see *f.limit.d* and *f.limit.r* in Appendix 2) are calculated using `free()` and the interesting regions are found by calculating the set minus of reference and patient free regions. The interesting regions of homozygosity are stored into *homozygousRegions* list.

The distribution of the total length of homozygous regions in patient samples is tested against the null hypothesis that it follows the normal distribution. `TestNormality()` function performs a Shapiro-Wilk normality test [SW65] and returns the p-value that is shown on the result graph.

The analysis of alleles is based on a reference vector (variable *col* in Figure 11) that divides alleles a and b into two haplotypes based on their frequencies:

$$col_c = \begin{cases} \underline{a}, & \text{if } frequency(\underline{aa}, \mathbf{S}_c) \geq frequency(\underline{bb}, \mathbf{S}_c) \\ \underline{b}, & \text{else.} \end{cases} \quad (9)$$

The vector of the most common alleles is produced in function `SetAlleleColors()` and it is used by `AlleleRegions()` to tell whether a SNP represents the common haplotype. `AlleleRegions()` compares *col* vector against each corresponding SNP of the given homozygous regions. The resulting set of regions contains the same homozygous regions that were given to the function but they have been split into haplotypes `red` and `blue`. The `red` haplotype represents the regions where the sample genotype $\mathbf{S}_c$ is homozygous to allele col_c and the `blue` refers to the regions of complement homozygosity. Heterozygous SNPs (the gaps), improperly hybridised SNPs, and SNPs that are not listed in *col* vector are ignored. The base pair division between the haplotype boundaries favours the `blue` regions so that the rare allele gets expanded to the limiting SNPs. The division policy maximises the sensitivity of detecting allele differences because the `blue` haplotype is unusual in references.

`CompoundRegions()` lists those regions that could be possible compound heterozygotes of the homozygous regions. The concept of the possible compound heterozygote and the detection algorithm are described in Section 4.4.

6.2.4 RegionAnnotator

RegionAnnotator is a Java program that can be used to collect biological information about the interesting regions found by CohortComparator. RegionAnnotator is capable of reading the result lists of CohortComparator and it can be given a score limit that tells which regions are worth a further analysis. Each interesting region is queried against Ensembl database for its biological content. The query interface is based on publicly

chromosome	start	end	score	type	samples
14	75683432	76374502	888394	1	C705,C93
14	93952397	94431793	875499	1	C383,C705
2	45292467	46457469	640784	1	C828,C835,C88
2	34804935	34858768	631390	1	C135,C828,C897
2	36444241	36566423	602266	1	C697,C828
6	24461416	24898997	455349	3	C69,C705,C828
1	43494738	44556103	454017	1	C279,C697,C702
5	102881726	103082033	443566	3	C279,C402,C770,C828,C897
13	53813508	54217913	401100	1	C222,C322,C383,C439,C791
13	62534529	62719498	364604	1	C322,C622
12	111494162	112203186	335760	1	C205,C222,C619
16	22705354	24061164	329410	1	C383,C442
⋮					

Figure 12: An example of the list of interesting regions. The file consists of rows of interesting regions sorted in descending order by the score. Each line starts with the location information given in base pairs followed by the actual score. The type of the region (1=region of homozygosity, 2=interesting allele, and 3=a joint region of the previous types) is given before a list of samples sharing the interesting feature.

available Ensembl API (Ensj) Java library that provides an object oriented view to the genomes and takes care of all object-relational mappings. Ensj is provided by the Ensembl hosts The Wellcome Trust Sanger Institute and European Bioinformatics Institute and it can be downloaded from the Ensembl web site[5].

The input format of RegionAnnotator is compatible with list of interesting regions produced by CohortComparator (an example is given in Figure 12) but only first three columns (chromosome number, start locus, and end locus) are mandatory. The scores can be used to limit annotation to the best scored regions if a score limit is given to RegionAnnotator. The input format of the program has been kept simple so that its use is not limited to CohortComparator but it can be used for general interests of querying transcripts encoded by the given parts of the genome.

[5]Ensembl genome browser can be found from http://www.ensembl.org/.

7 Results

The amount of SNP data used in this colorectal cancer study was so huge that all methods of analysis needed to be scalable. The final analysis was completed with 51 reference and 42 patient samples each consisting of 57290 and 112990 SNPs, respectively. The amount of data to be processed simultaneously was reduced by splitting haplotype matrices of both data sets into chromosome specific files.

The important questions to be answered were:

1. Which are the loci of homozygous regions shared by some patients but absent in references?
2. Which are the loci of homozygous regions with alleles absent in references?
3. Are there any homozygous deletions shared by some patients but absent in references?
4. Which are the samples sharing any of these features?
5. What are the genes in these regions?

CohortComparator and RegionAnnotator were written in order to answer these questions. The designing, construction, testing, optimisation and analysis of the applications are my most important contributions to this study. CohortComparator takes many parameters (see Appendix 2) that can be fine tuned based on what kind of experiments it is used for. Changing these parameters would be hard and it could lead to misleading results unless the meaning of the argument is well understood. The analyses of these arguments are given in sections 7.2 and 7.3.

7.1 Distribution of homozygous SNPs

The first step in comparing the data sets $\mathbf{S}^D$ and $\mathbf{S}^R$ was to check whether they have similar distributions of homozygosity in all SNPs. The high level comparison was done because it was not known how similar the distributions would be. The interesting regions could be detected without a complex analysis if the distributions would differ enough in certain regions.

Homozygosities were calculated for each SNP in both data sets and visualised in a graph. Rates of homozygous samples were compared between both data sets and for each SNP in order to see whether the differences were of any significance. Uncorrected p-values were smoothed with the same locally weighted linear regression (LOWESS function)

[Cle79] as the homozygosity levels. Only those p-values are drawn to the graph, which are less than 0.2. A pair of examples of homozygosity level comparison graphs can be

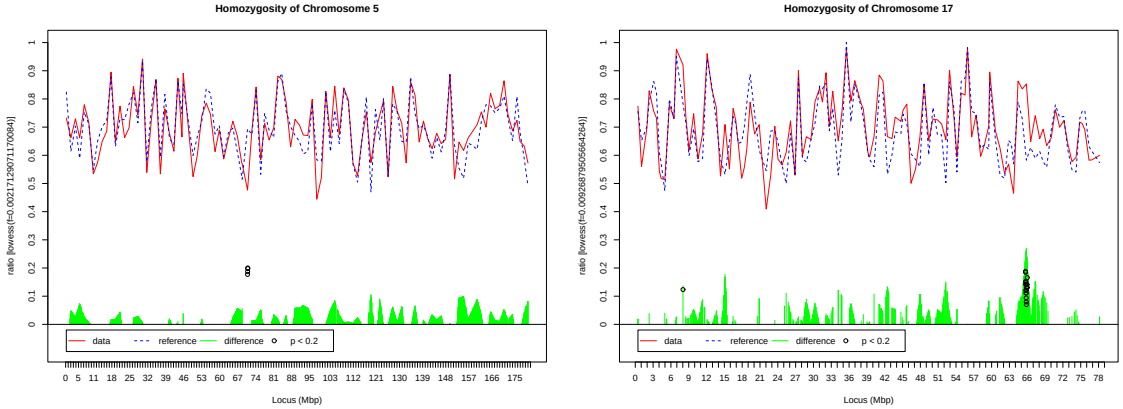

Figure 13: Two example graphs of homozygosity levels. The vertical axis represents the smoothed ratio of homozygous genotypes within the sample set. The chromosomal loci are shown on x-axis. The vertical green bars show the positive differences between the patient and reference sets i.e. the regions that have more homozygosity in patient data.

seen in Figure 13. Chromosome 5 is shown in here so that it can be compared against other figures given in this thesis (like those in Appendix 1). The graph of chromosome 17 has been chosen here as it has clear regions of smoothed p-values less than 0.2. The smoothing of the p-values is a visualisation trick to keep values comparable in the graph and to emphasize those regions that may harbour interesting differences. The statistical relevance of such measures is, however, not clear. The application of a real false discovery rate correction [PC06] revealed no significant findings in any chromosome. The result was expected because the sample size was relative small comparing to number of SNPs ($n \ll k$) and because of the big variance in p-values. The statistical comparison of allele distributions of each SNP was not sensitive enough to detect weak signals of continuous homozygous regions as there were other homozygous SNPs (fractions with length $< l$) confusing the test.

7.2 Determination of a significant region length

The chance of getting homozygous region by chance was analysed using randomly generated samples in linkage equilibrium. The frequencies of aa, ab, and bb were calculated from each cancer patient sample and 51 random samples of chromosome 5 were generated. The SNP specific genotype frequencies were used to weight the genotype probabilities and each SNP was assigned a value independently. Both, number of homozygous

regions and their chromosome coverage, were calculated with different values of l and the same analysis was completed with the original cancer patient data. The results are given in Figure 14. The number of homozygous regions found from the random samples decreased well below 0.05, when $l = 54$ but an average of ~ 4.67 homozygous regions were still found from the patient samples. The SNP genotypes of the real samples were clearly dependent on surrounding SNPs and distinctive from the independent genotypes of random data.

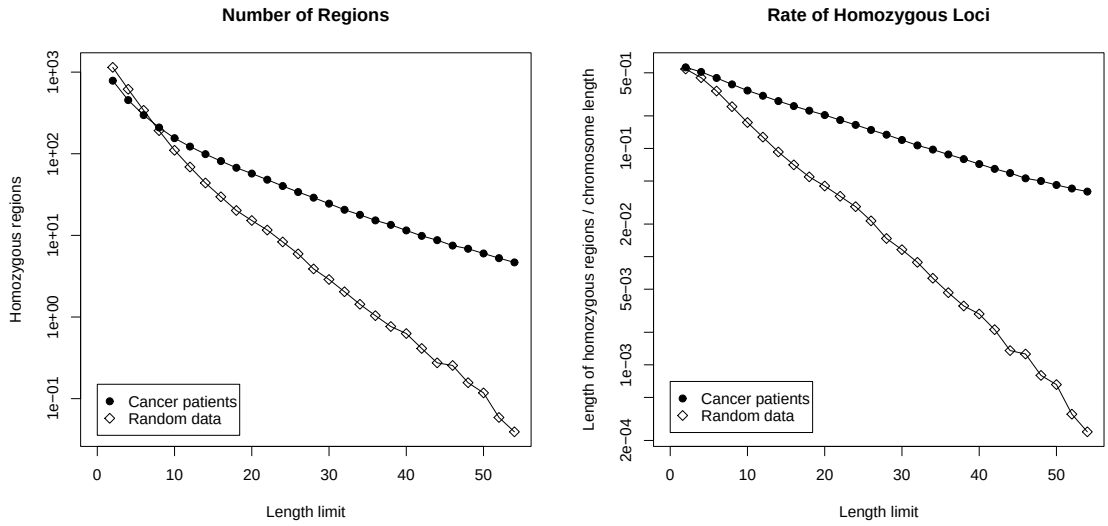

Figure 14: The distribution of homozygous regions of cancer patients in comparison to random data. The y-axes of both figures are in logarithmic scale and the values are normalised in according to the number of samples analysed.

An increase in minimum length of the homozygous region (l) of interest decreases the number of regions we expect to see but the chance of getting such regions without a biological relationship decreases. Biological relevance of the findings is important but on the other hand we should lose as little information as possible. The balance between these two exclusive objectives is challenging to answer analytically because the biological relevance cannot be determined a priori. Long regions are more likely to be autozygous [Woo04] but the mutations we are interested in may be on shorter fragments. Only some estimates of the evolutional background of the haplotypes can be done (see Section 4.4) since we do not know the pre-existed genomes nor complete pedigrees. Figure 16 illustrates the synergism of l parameter, when adjusted simultaneously in both patient data and reference data (parameters *l.limit.d* and *l.limit.r* in the figure). Figure 15 illustrates

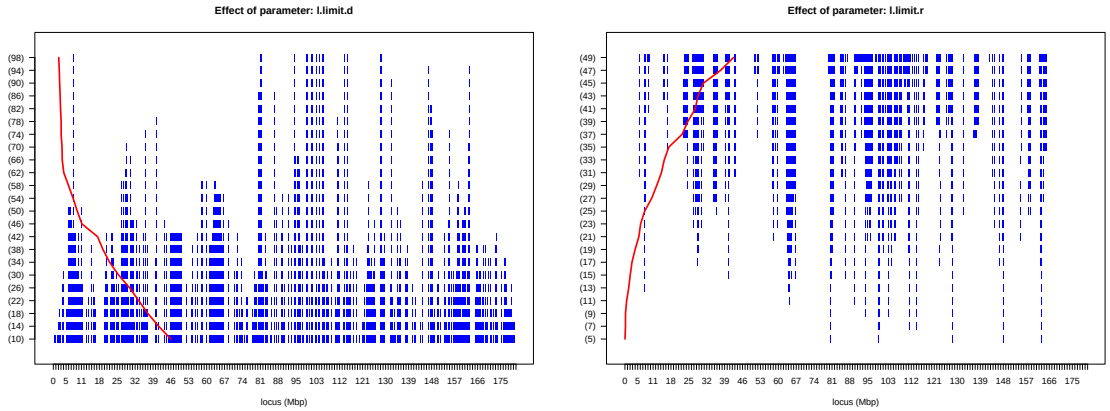

Figure 15: Total amount of interesting regions in comparison to l^D (on the left) and l^R (on the right). In both graphs, the length limit of the other cohort is kept constant ($l^R = 25$, $l^D = 55$). The value of the length limit is given on the y-axis (in numbers of SNPs). The red curve represents the total length of the interesting regions (blue bars) and it follows the scale of the x-axis. The same axis has been used for the physical locations of the regions but in that case the base pairs should be considered as chromosome loci.

the separated examination of both parameters [6]. The horizontal bars represent interesting regions found in chromosome 5. The length of the bar does not reflect the length of individual homozygous regions of the samples but it consists of at least two overlaps of such features (homozygosity or allele) that are not present in references. The definition of interesting region is further discussed in Section 4.3. Numbers of interesting regions decreases if reference set criteria are relaxed or they are tightened in data set. The balance between the two exclusive length limits is met around the values of $l^D = 54$, $l^R = 27$.

The clear imbalance in accuracy between the patient (HindIII and XbaI chips used) and reference (no HindIII chip data) sets lead to difficulties in parameter determination in order to keep both sets comparable. Region length limit (l^R) for the reference set was estimated so that the physical length of the regions found became the same in both data sets: $min(length(\mathbf{R}^D)) = min(length(\mathbf{R}^R))$. The effect of data set differences was eliminated by using XbaI SNPs of the cancer patient samples as references. l^D was set to 55 and the analysis was conducted using different values of l^R. The lengths of the shortest regions of both cohorts were compared by eye and they were found to be almost equal when $l^R = 25$. Results of this experiment can be seen in Figure 24 in Appendix 1.

[6] Figures 16 and 15 were generated using gap distance method with parameters: $d = 8$, $f.limit.d = 0.03$, and $f.limit.r = 0$. Constants used in Figure 15 were $l.limit.d = 55$ and $l.limit.r = 25$.

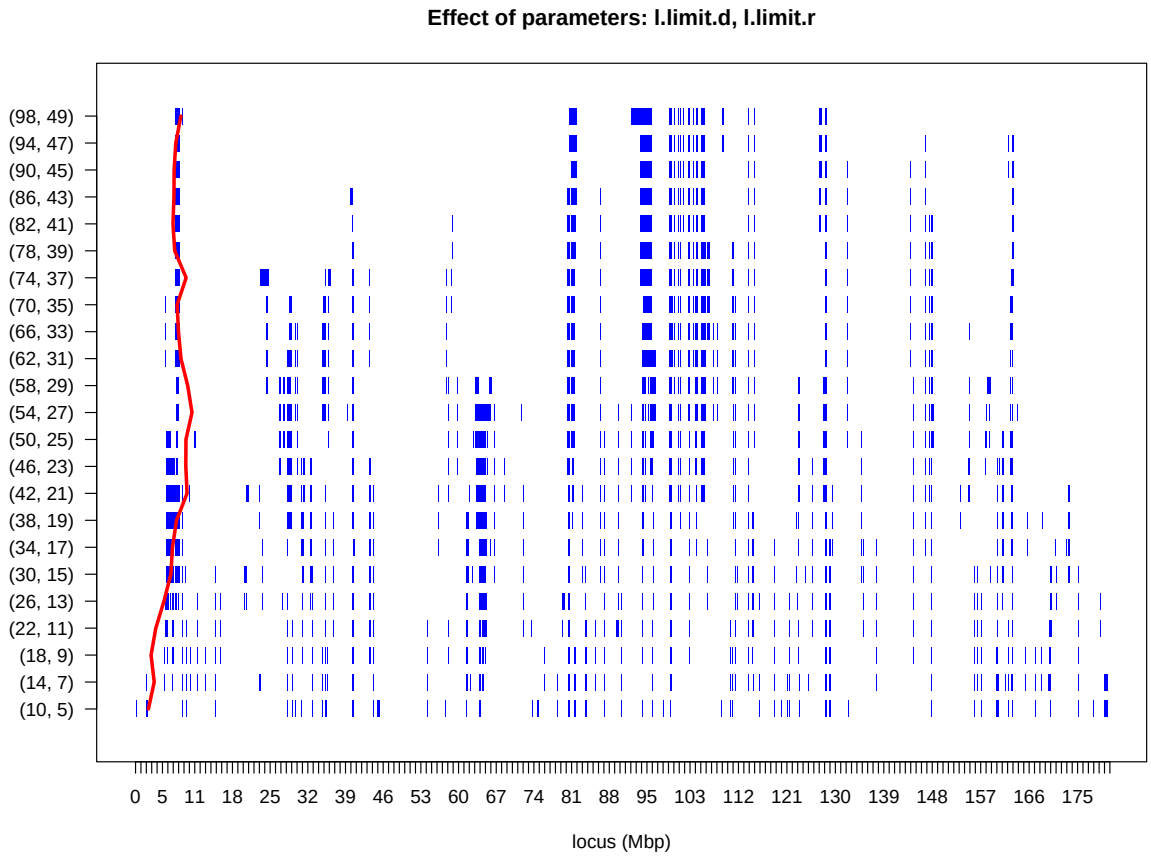

Figure 16: The compound effect of l parameter. Y-axis consists of pairs of length limits (l^D, l^R). The red curve depicts the total length of interesting regions.

7.3 Properties of the region detection algorithms

The window length parameter w used in sliding window method is not critical for the results. Combinations of different window lengths for reference and patient data show only minor differences when $1/g \leq w \leq l$. The lower bound of $w \geq 1/g$ allows at least one heterozygous SNP within the window whereas the gap acceptance benefits of the method are lost with the shorter windows (each heterozygous SNP interrupts the homozygous region). Window sizes greater than l are somewhat irrelevant since we are interested in shorter homozygous regions surrounded by heterozygous continuums. A too long window would not allow gap in such islands because the g would be exhausted by the surroundings. An example set of experiments on chromosome 5 is given in Figure 17. There are two components in gap ratio as defined in "window" method of CohortComparator. The first component tells the window length (w) and the second one gives the maximum ratio of heterozygous SNPs (g) within the window. The second component is kept constant

(0.1) in these experiments and omitted from the figure. The minimum length of the regions of interest (l) is 25 for the reference set and 55 for the patient samples [7]. There are

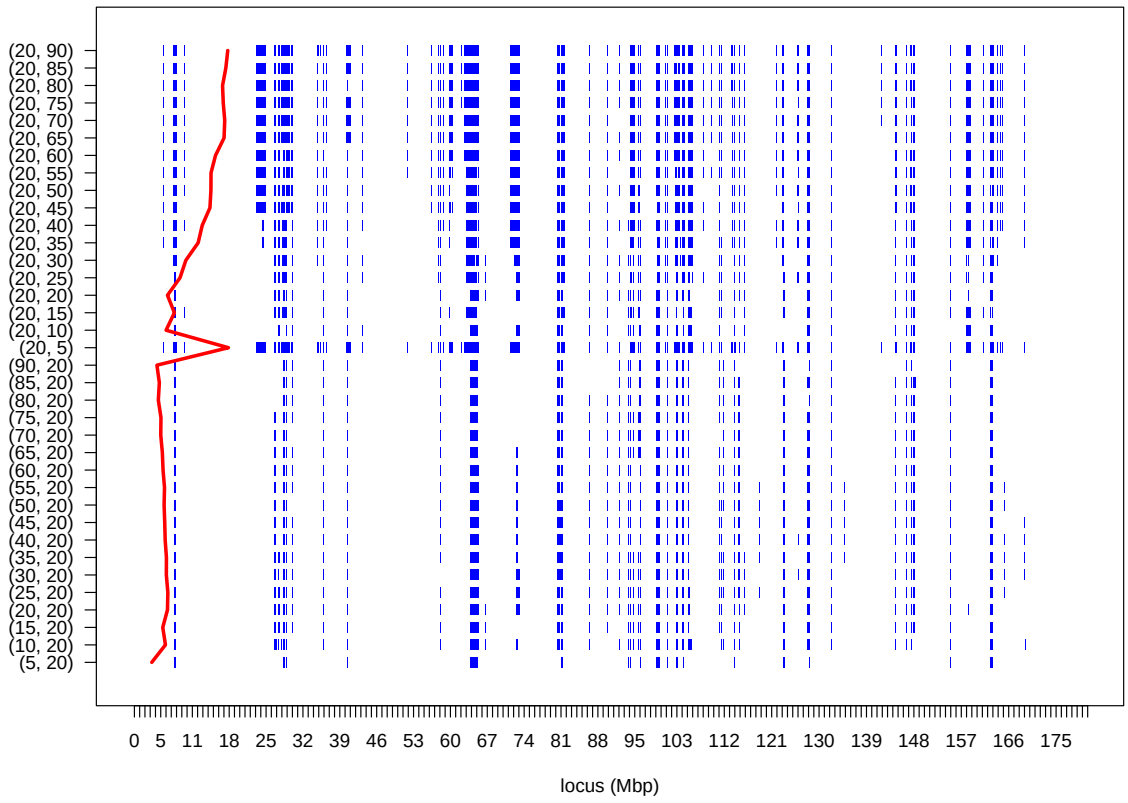

Figure 17: Effects of the length of the sliding window in respect to the interesting regions found. The window lengths for the patient samples and the references are given on y-axis as tuples of (w^D, w^R). The window lengths are measured as a number of SNPs. Total length of the interesting regions is shown with a red curve.

not many differences between those rows in Figure 17 that fit into range (10–55, 10–25) and even the rows outside of the range are fairly similar. The considerably big leap from row (20, 5) to row (20, 10) can be explained by the fact that no heterozygous gaps are allowed when the window length is less than 10 whereas at least one heterozygous SNP fits into the range when $w \geq 10$. Saturated patterns outside of the range are produced by the gapless homozygous regions. The pure regions are rare comparing to those with gaps and we observe more regions when the parameter saturates in reference data (less filtering) and only few short regions in case of patient data (less region candidates).

The use of gap distance method means the substitution of parameters w and g by the gap

[7] Other parameters used in Figure 17 were: $f.limit.d = 0.03$, and $f.limit.r = 0$.

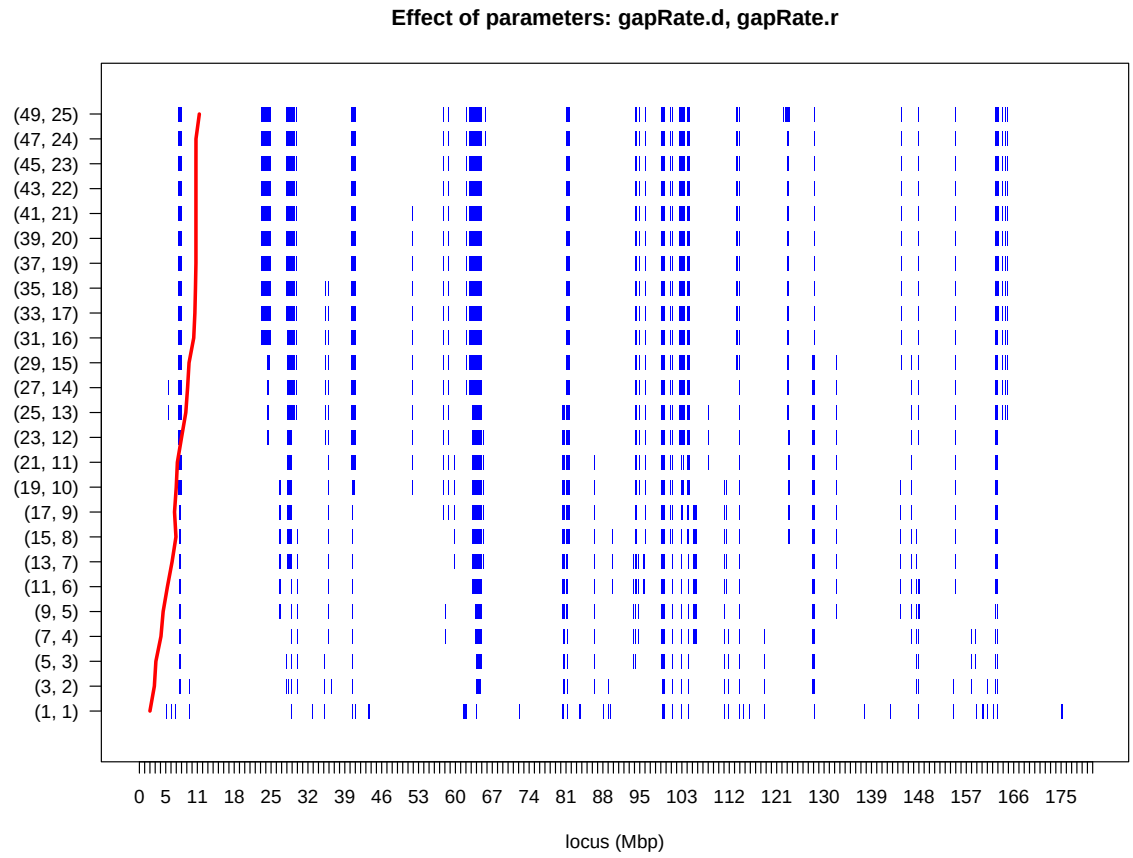

Figure 18: Minimum distance between heterozygous SNPs (d^D, d^R). The blue bars represent interesting regions found from the chromosome 5 using the gap distance method with different values of d^D, d^R as shown on the left. The other parameters used in this example are: $s = 1$, $f.limit.d = 0.03$, $f.limit.r = 0$. The total length of the interesting regions is illustrated by the red curve.

distance measure d. The average distance between the gaps can be given as $1/g - 1$ but the method gives no compensation to gaps with distances below the d although they would have been surrounded by long and gapless sequences. The ignorance of overall gap rate means that $d \leq 1/g - 1$. The difference between the$d$ and g measures depends on the distribution of the gaps, where the equality holds for the uniform distribution only.

It is not surprising that the changes in d seem to follow the same pattern as changes in g as there is a close relationship between these two parameters. The effects of the parameter to the list of interesting regions are illustrated in Figure 18. The total length of the interesting regions increases as the distance between the gaps becomes longer. Reason for this phenomenon can be found from the amount of homozygous regions within the reference set. Features of the patient data are less likely to be ignored as the amount of

reference regions gets smaller. The saturation of the regions occurs when the d^D becomes greater than 25 and d^R becomes greater than $d^D/2$. The regions found after the saturation are almost solely determined by the parameters l and s.

An optional support for base pair metrics in SNP distances has been implemented into CohortComparator as an alternative to counts of intermediate SNPs. The base pair distances can be used in order to compensate the effects of the non-uniform SNP distribution. The length of the sliding window and l were both calculated in base pairs when `bpLength` option was turned on. The advantage of this approach is that the parameters are automatically calibrated to both data sets with different SNP densities (see Section 6.1) but the probability of getting false positives in areas with low SNP densities was increased as the amount of SNPs associated to each region varied. Figure 23 illustrates how the low density areas close to telomeres (both ends) and the centromere (46–50Mbp) contain homozygous regions not present in Figure 21, which was created using SNP limits $l^R = 25$ and $l^D = 55$. The DNA sequences of telomeres and centromeres are typically quite conserved, which advocates real regions but the true existence of mutations remains obscure unless the samples are sequenced.

Gap distance, gap ratio, and sliding window methods are all equal (to **R** in Equation 2) if no gaps are allowed. The behaviour of these methods has been empirically compared using realistic values of parameters l, s, d, g, and w as suggested by the analysis explained in this section. An example set of results of all three methods is given in Figure 19. It can be seen that the interesting regions are found from the same areas but the exact loci varies between the algorithms. All algorithms agree on 34 regions but the variation around these regions is enough to change the list of affected genes. The analysis of the scores of the interesting regions was performed with the same method as in Figure 7. The method is not as accurate with one chromosome as it is with a genome wide data but the limit of 11800 was chosen based on the suggested values of 14313 (gap distance), 12827 (sliding window), and 11826 (gap ratio). Comparison of the regions exceeding the limit revealed that all 5 regions found by the gap distance method were also found by the other two methods. The results of gap ratio method included 6 regions that were not in top 5 given by the gap distance method. Only one of the six regions returned by the gap ratio method was also detected by the sliding window. Sliding window had 3 interesting regions with scores greater than 11800 not shown in the list of gap distance method. Top best scoring regions were considered to be the same if they overlapped but differences in the limit of the regions were observed in regular basis. The comparison of the sample lists associated to each region revealed cases were a sample was associated to the region by one algorithm but left out by another one.

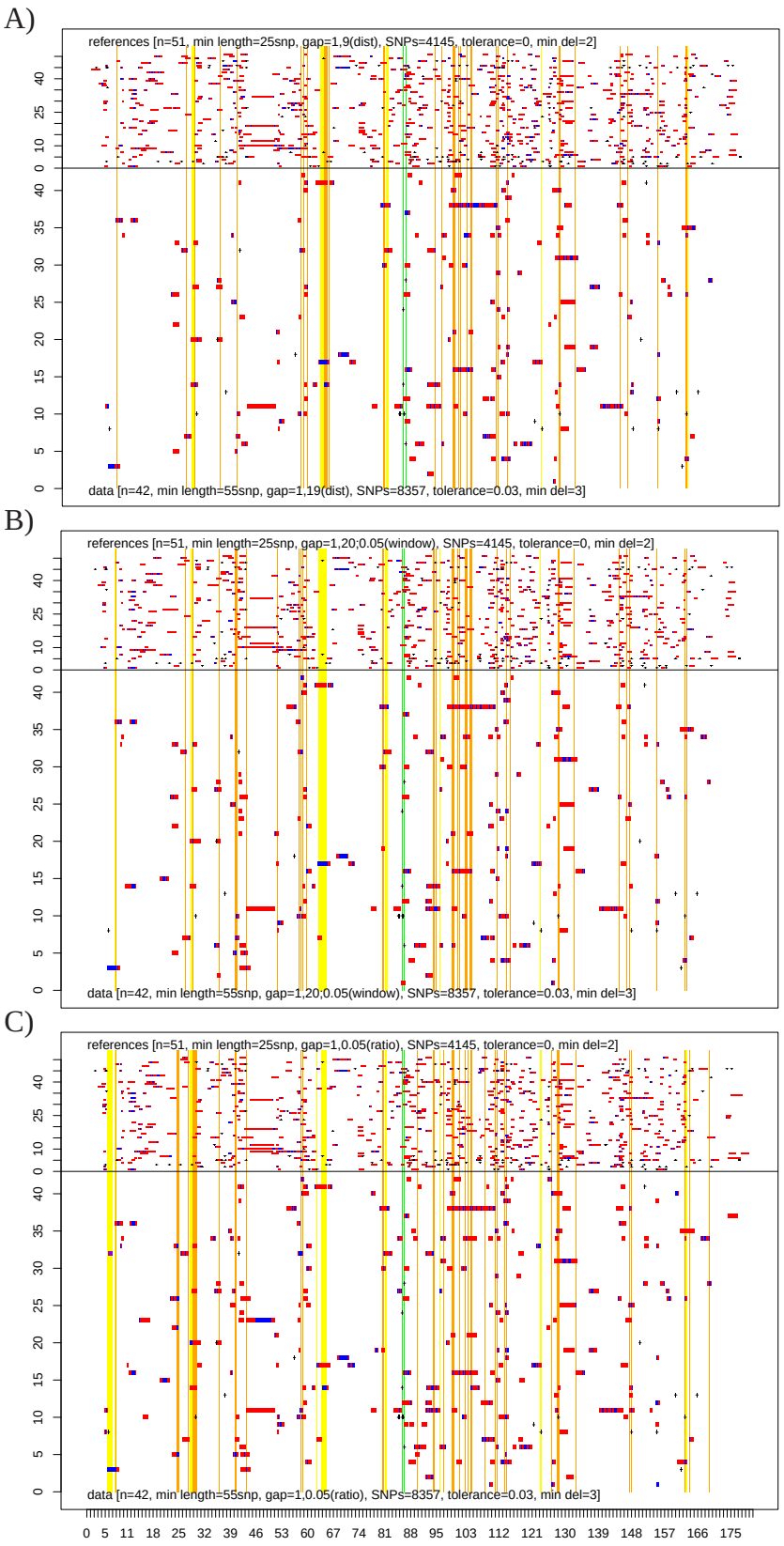

Figure 19: Homozygous regions as found by A) the gap distance, B) the sliding window, and C) the gap ratio methods. This example study of the chromosome 5 illustrates the small variation between the definitions of the homozygous regions. A detailed explanation of the formalism used in these graphs is given in Appendix 1.

7.4 Mutation detection probability

A criterion to measure the usefulness of a candidate region detection method is its ability to detect mutation sites and to reduce the length of genome under surveillance. In this section we describe how likely our method labels a mutation site as a candidate region.

The probability of considering a region to be interesting (P_{int}) is a joint probability of having at least t^D overlapping homozygous regions (P^D) in case cohort and a probability of no more than t^R overlapping regions in references ($1 - P^R$):

$$P_{int} = P^D \times (1 - P^R). \tag{10}$$

Let h^D be the probability of observing a homozygous region at given locus. The probability h^D is defined for the sample space of patients and it should[8] represent relative frequency of homozygous regions at the locus. The probability of at least t^D overlapping regions can be calculated using the probability mass function of binomial distribution:

$$P^D(h^D, n^D, t^D) = 1 - \sum_{i=0}^{t^D-1} \binom{n^D}{i} (h^D)^i (1 - h^D)^{n^D - i} \qquad t^D \in [1, n^D]. \tag{11}$$

A similar approach can be applied to references. First, h^R is defined as the probability of having a homozygous region in one sample at a certain locus. Secondly, the probability of at least t^R overlapping regions is determined:

$$P^R(h^R, n^R, t^R) = \begin{cases} 1 - \sum_{i=0}^{t^R-1} \binom{n^R}{i} (h^R)^i (1 - h^R)^{n^R - i}, & t^R \in [1, n^R], \\ 1 - (1 - h^R)^{n^R}, & t^R = 0. \end{cases} \tag{12}$$

Equation 12 has been defined separately for $t^R = 0$ to avoid negative indices in sum function but otherwise it has been derived from the same binomial probability function as Equation 11.

In comparison of homozygous regions we assume that $P^D > P^R$ for those regions that contain mutations significant to observed phenotype. The reasoning for this assumption is in the definition of the case sample space that suggests a recessive phenotype. Controls are less likely to share the same homozygous haplotype as they express a different phenotype.

Limits of t^D and t^R represent prior hypotheses of how many overlapping homozygous regions are expected in the cohorts. The values of these limits could be estimated using their expectation values if h^D and h^R are known: $t^D = n^D \times h^D$ and $t^R = n^R \times h^R$. Determination of realistic limits provides a reasonable balance between false positives and sensitivity.

[8] Assuming that the detections of homozygous regions causes no false positives or false negatives.

Obviously, P_{int} can be maximised by using values $t^D = 1$ and $t^R = n^R$ but that would make no use of reference samples and all regions with at least one homozygous sample would be classified interesting. The strengthening of the limits decreases the number of possibly interesting loci. On the other hand, it becomes more likely that the mutation region contains too few overlapping homozygous regions in case samples or there would be too many mutant haplotypes in reference set.

In this simulation, values of P_{int} have been estimated using Equation 10. Simulations have been conducted using various cohort sizes (assuming $n^D = n^R$), where h^D has been fixed to 0.05 and h^R has varied between 0 and 0.05. Examples of these simulations are shown in Figure 20. Each graph represents P_{int} estimates (red lines) for one n with two alternative h^D (0.05 and 0.1). The fluctuation of the lines is caused by the discrete expectation value estimates of t^R; the sensitivity of the method increases as more reference regions are tolerated. Statistical difference between the cohorts is measured using Kolmogorov-Smirnov test (blue lines) [Pre02] with a null hypothesis that $t^D = t^R$. The shown probability values for the null hypothesis have been calculated as a mean of tests applied to 100 random cohort pairs. The cohorts were generated using values of h^D and h^R.

It can be seen in Figure 20 that the comparison of homozygous regions is suitable for small cohorts ($n < 100$) but does not benefit from the additional information provided by larger cohorts. The statistical analysis of distributions may be more practical if there are a vast number of samples available. The differences between the distributions of homozygous regions between the cohorts are so minor that the statistical significance remains low even if $n = 1000$. P_{int} decreases rapidly in function of h^R if the change is not compensated by the increase in t^R. The observation is quite obvious as the increase in t^R increases the change of having mutation haplotypes in reference samples. The changes in h^D are well tolerated as $\lim_{n^D \to \infty} P^D = 1$, when $h^D < 1$. This means that a region is considered to be interesting if the amount of overlapping regions exceeds limit t^D and no difference is made on how far this limit was exceeded.

7.5 Effects of the haplotyping

The amount of homozygous regions found in the reference set is higher than in experiment data with HindIII SNPs included because less homozygous SNPs are needed and such regions may be caused my non-autozygous regions of (almost) monoalleleic markers. The dramatic effect of false positives caused by the sparse SNP set can be seen by comparing Figure 24 and Figure 25. Both figures have been created using the patient data but the

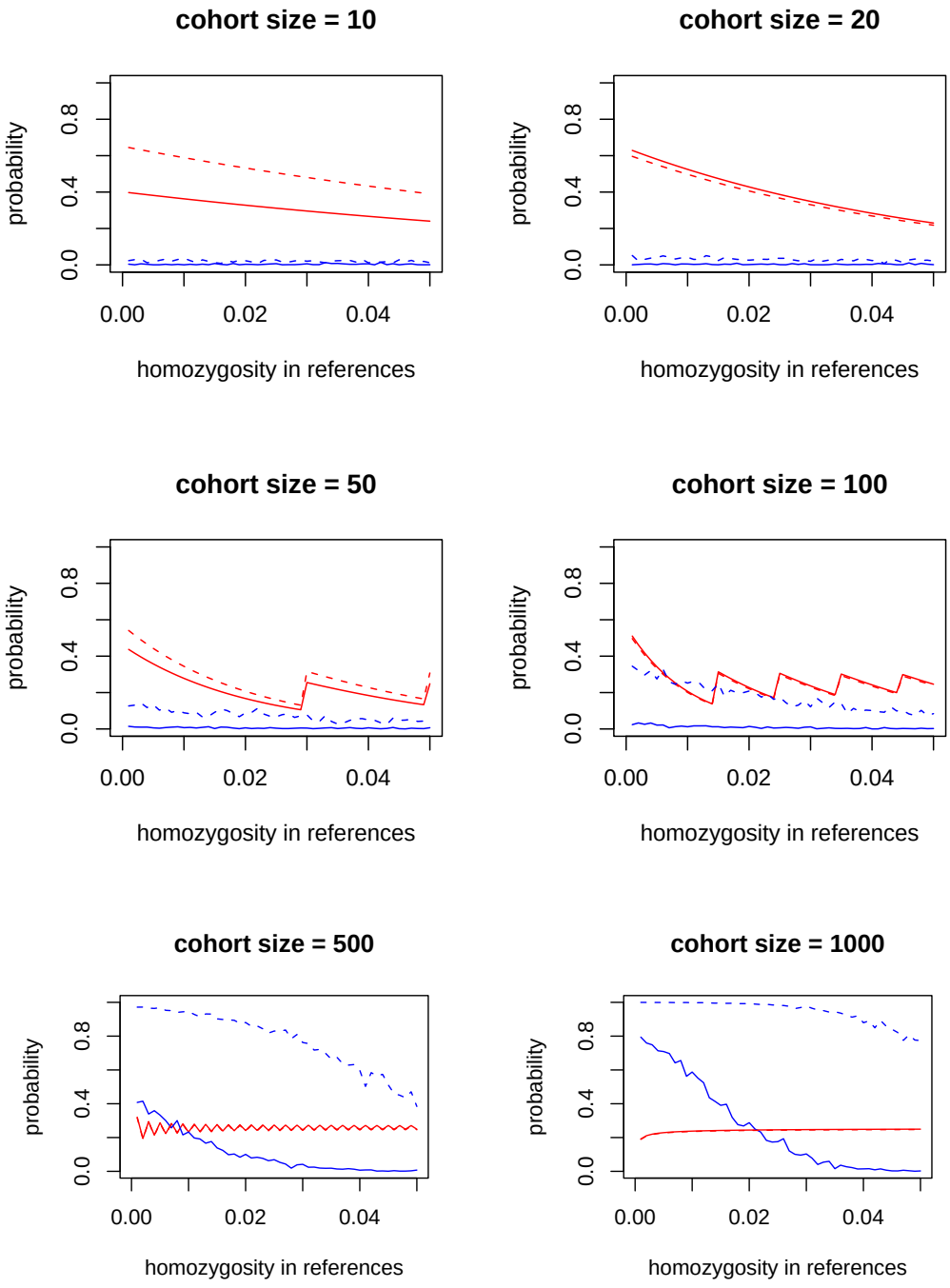

Figure 20: Simulated values of mutation detection probabilities. Red lines represent a probability of labelling a mutation to an interesting region. The chance of a homozygous region at the mutation site has been set to 1/20 (solid lines) and 1/10 (dashed lines). The same probability for the reference set is shown on the x-axis. Kolmogorov-Smirnov test probabilities are shown in blue.

SNP resolution differs between the data sets. The first figure illustrates the case where the reference set lacks the HindIII information. The lower SNP resolution leads to a greater amount of homozygous regions. No interesting regions are expected in these figures as the samples are compared to themselves. However, the second figure contains interesting regions all over the chromosome because the dense set of homozygous regions found from the sparse SNP set is not compensated by the homozygous regions found from the HindIII and XbaI data. The resolution of 10^5 SNPs may sound extremely good but it seems that our method suffers from having fewer markers for a genome wide analysis. Further, it has been reported that a genome wide haplotype analysis may require as many as 10^6 SNPs [HP04]. The vast amount of questionable homozygous regions found from the reference data was likely to mask some real mutations.

The haplotype estimation of missing HindIII values was tried in order to increase l^R and to reduce the false positives that would not meet the selection criteria if measured with XbaI chip. The estimation was conducted by using fastPHASE haplotype prediction software. The haplotypes of the population were calculated from all samples ($\mathbf{S}^D$ and $\mathbf{S}^R$) and the HindIII SNPs of the references were set to missing values. All information about the HindIII alleles comes from the patient data, which may lead to a bias towards patient haplotypes. The mutation in interest is probably shared by only few individuals (see Section 6) where the rest of the samples are "normal". The alleles of the normal samples are more likely to bind the references as they share the common haplotype whereas the mutation is likely to differ in its XbaI constitution also. The computationally predicted values were converted back to SNP data matrices with the full SNP resolution and the analysis was conducted using the original patient data parameters for both data sets.

The results of the haplotype estimations were in better balance than the original results, what comes to the spatial densities of homozygous regions along the chromosomes. An example of the results in chromosome 5 is given in Figure 22 and the same chromosome is shown without haplotype estimations in Figure 21. The chromosomal areas of interesting regions are quite well conserved between the original data and the haplotype estimates but the short fragments of the original data have fused into wider regions. The genome wide analysis of the original data returned 767 interesting regions covering total of 70.868960 Mbp. The estimation of the haplotypes decreased the number homozygous regions in reference data, which led to 1940 interesting regions with total of length of 199.752217 Mbp. The average length of an interesting region is increased[9] by 10557.5 bp, which is over 28 times more than the change in score limit that is calculated as described in Figure

[9] from 92397.6 to 102965.1 bp

		count	minimum	median	mean	maximum	$10^{12}\sigma^2$
references	before	7794	97770	1079000	1395000	47270000	2.230
	after	3831	280200	1211000	1530000	29030000	1.788
	NoCalls	3680	280200	1206000	1518000	29030000	1.660
patients	before	2428	262800	1232000	1676000	34580000	3.108
	after	2638	262800	1244000	1685000	34580000	3.239
all samples	before	10222	97770	1121000	1462000	47270000	2.452
	after	6469	262800	1226000	1593000	34580000	2.385
	NoCalls	6118	262800	1216000	1581000	34580000	2.240

Table 5: Length distribution of the homozygous regions before and after the haplotype estimation of the missing data. NoCalls rows represent cases, where the haplotype estimated data was populated with `NoCall` values copied from the original XbaI data. Genome wide counts of the homozygous regions are given on the first column. The region lengths are given in base pairs and the variance (σ^2) is shown on the last column.

7. The original score limit of 12987 was replaced with 13357 due the estimation of the haplotypes. The amount of regions exceeding the limit was increased from 63 to 165 regions.

The lengths of the homozygous regions and their total number were counter before and after the haplotype estimation. The preceding values are summarised in Table 5. The effects of the haplotype estimations into region lengths are first listed separately for both data sets and next together to illustrate their synergy. More than half of the reference values of the last data set consist of estimated values origin from the patient data. An extensive decrease in homozygous regions would be seen if the estimation was a random process as the probability of getting a sequence of 55 homozygous SNPs is seemingly low. The median length of the regions was increasing close to that of cancer patients, although the number of homozygous regions was almost halved. Taking this into consideration there is a reason to believe that the number of regions was not reduced that much because of a poor estimation algorithm but because of the changes in length limit.

7.6 Biological content of the candidate regions

Annotations of the all 1940 interesting regions revealed 2041 genes classified as novel or known. The results of the comparison between the genes and the colorectal cancer genes given in SNPs3D are listed in Table 6. The association between the SNPs3D gene name and the region annotation is based on a simple text search. The order of the genes reflects

the association strength between the gene and the colon cancer as given by SNPs3D.

The original gene annotation set was also compared against the CRC genes listed in [Sjö06] and only five genes were observed to overlap between these two lists and *GFRA1* was associated to *RET* based on its description. The list of genes revealed by the text search is given in Table 7. *GFRA1* was the only gene found from both tables 6 and 7.

SNDs3D ref.	locus	gene name	description
MLH3	14: 74553239-74587886	MLH3	DNA mismatch repair protein Mlh3 (MutL protein homolog 3).
ATM	11: 107598769-107745036	ATM	Serine-protein kinase ATM (EC 2.7.11.1) (Ataxia telangiectasia mutated) (A-T, mutated).
FHIT	3: 59712992-60497735	FHIT	Bis(5'-adenosyl)-triphosphatase (EC 3.6.1.29) (Diadenosine 5',5"'- P1,P3-triphosphate hydrolase) (Dinucleosidetriphosphatase) (AP3A hydrolase) (AP3AASE) (Fragile histidine triad protein).
ERBB2	17: 35104766-35138441	ERBB2	Receptor tyrosine-protein kinase erbB-2 precursor (EC 2.7.10.1) (p185erbB2) (C-erbB-2) (NEU proto-oncogene) (Tyrosine kinase-type cell surface receptor HER2) (MLN 19).
DPYD	1: 97315887-98159193	DPYD	Dihydropyrimidine dehydrogenase [NADP+] precursor (EC 1.3.1.2) (DPD) (DHPDHase) (Dihydrouracil dehydrogenase) (Dihydrothymine dehydrogenase).
NFKB1	4: 103641518-103757506	NFKB1	Nuclear factor NF-kappa-B p105 subunit (DNA-binding factor KBF1) (EBP- 1) [Contains: Nuclear factor NF-kappa-B p50 subunit].
MAP2K4	17: 11864866-11987865	MAP2K4	Dual specificity mitogen-activated protein kinase kinase 4 (EC 2.7.12.2) (MAP kinase kinase 4) (JNK-activating kinase 1) (c-Jun N-terminal kinase kinase 1) (JNKK) (SAPK/ERK kinase 1) (SEK1).
TNF	6: 24758184-24775240	TTRAP	TRAF and TNF receptor-associated protein (ETS1-associated protein 2) (ETS1-associated protein II) (EAPII).
SELP	1: 167824661-167866023	SELP	P-selectin precursor (Granule membrane protein 140) (GMP-140) (PADGEM) (Leukocyte-endothelial cell adhesion molecule 3) (LECAM3) (CD62P antigen).
CGA	6: 87851935-87861569	CGA	Glycoprotein hormones alpha chain precursor (Anterior pituitary glycoprotein hormones common subunit alpha) (Follitropin alpha chain) (Follicle-stimulating hormone alpha chain) (FSH-alpha) (Lutropin alpha chain) (Luteinizing hormone alpha chain)
GATA4	8: 11599122-11654920	GATA4	Transcription factor GATA-4 (GATA-binding factor 4).
PEX7	6: 137185410-137276752	PEX7	Peroxisomal targeting signal 2 receptor (PTS2 receptor) (Peroxin-7).
PLA2G4A	1: 185064708-185224736	PLA2G4A	Cytosolic phospholipase A2 (cPLA2) (Phospholipase A2 group IVA) [Includes: Phospholipase A2 (EC 3.1.1.4) (Phosphatidylcholine 2- acylhydrolase); Lysophospholipase (EC 3.1.1.5)].
DLC1	8: 12985243-13416766	DLC1	Rho-GTPase-activating protein 7 (Rho-type GTPase-activating protein 7) (Deleted in liver cancer 1 protein) (Dlc-1) (HP protein) (StAR-related lipid transfer protein 12) (StARD12) (START domain-containing protein 12).
WNT5A	3: 55479112-55489996	WNT5A	Protein Wnt-5a precursor.
FAP	2: 162735446-162808291	FAP	Seprase (EC 3.4.21.-) (Fibroblast activation protein alpha) (Integral membrane serine protease) (170 kDa melanoma membrane-bound gelatinase).
HMGB1	13: 29930884-30089729	HMGB1	High mobility group protein B1 (High mobility group protein 1) (HMG- 1).
SRC	3: 36397101-36564500	STAC	SH3 and cysteine-rich domain-containing protein (SRC homology 3 and cysteine-rich domain protein).
SMARCA3	3: 150230604-150287007	SMARCA3	SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily A member 3 (EC 3.6.1.-) (Sucrose nonfermenting protein 2-like 3) (DNA-binding protein/plasminogen activator inhibitor 1 regulator) (Helicase-like transcription factor)
RAD51	14: 67360014-68187315	RAD51L1	DNA repair protein RAD51 homolog 2 (R51H2) (RAD51-like protein 1) (Rad51B).
SELE	1: 167958406-167969827	SELE	E-selectin precursor (Endothelial leukocyte adhesion molecule 1) (ELAM-1) (Leukocyte-endothelial cell adhesion molecule 2) (LECAM2) (CD62E antigen).
MAPRE2	18: 30875379-30975820	MAPRE2	Microtubule-associated protein RP/EB family member 2 (APC-binding protein EB2) (End-binding protein 2) (EB2).
PHC1	12: 8959623-8985325	PHC1	Polyhomeotic-like protein 1 (hPH1) (Early development regulatory protein 1).
DVL3	3: 185355978-185374092	DVL3	Segment polarity protein dishevelled homolog DVL-3 (Dishevelled-3) (DSH homolog 3).
RING1	6: 33284255-33288477	RING1	Polycomb complex protein RING1 (RING finger protein 1).
MYB	6: 135544146-135582004	MYB	Myb proto-oncogene protein (C-myb).
BDKRB1	14: 95799760-95800847	BDKRB1	B1 bradykinin receptor (BK-1 receptor) (B1R).
BDKRB2	14: 95740950-95780536	BDKRB2	B2 bradykinin receptor (BK-2 receptor) (B2R).
CCNG1	5: 162797155-162804598	CCNG1	Cyclin-G1 (Cyclin-G).
PLXDC1	17: 34473083-34561298	PLXDC1	Plexin domain-containing protein 1 precursor (Tumor endothelial marker 7) (Tumor endothelial marker 3).
ANTXR1	2: 69093993-69329804	ANTXR1	Anthrax toxin receptor 1 precursor (Tumor endothelial marker 8).
VCP	9: 35046061-35063246	VCP	Transitional endoplasmic reticulum ATPase (TER ATPase) (15S Mg(2+)- ATPase p97 subunit) (Valosin-containing protein) (VCP).
RPL31	2: 100985183-100989312	RPL31	60S ribosomal protein L31.
TGFA	2: 70527927-70634438	TGFA	Transforming growth factor alpha precursor (TGF-alpha) (EGF-like TGF) (ETGF) (TGF type 1).
CFTR	7: 116907253-117095951	CFTR	Cystic fibrosis transmembrane conductance regulator (CFTR) (cAMP- dependent chloride channel) (ATP-binding cassette transporter sub- family C member 7).
PPP1R1B	17: 35036705-35046403	PPP1R1B	Dopamine- and cAMP-regulated neuronal phosphoprotein (DARPP-32).
CYP2C9	10: 96688418-96739137	CYP2C9	Cytochrome P450 2C9 (EC 1.14.13.80) ((R)-limonene 6-monooxygenase) (EC 1.14.13.48) ((S)-limonene 6-monooxygenase) (EC 1.14.13.49) ((S)- limonene 7-monooxygenase) (CYPIIC9) (P450 PB-1) (P450 MP-4/MP-8) (S-mephenytoin 4-hydroxylase) (P-450MP).
ALOX5	10: 45189635-45261567	ALOX5	Arachidonate 5-lipoxygenase (EC 1.13.11.34) (5-lipoxygenase) (5-LO).
PLXDC2	10: 20145174-20609292	PLXDC2	Plexin domain-containing protein 2 precursor (Tumor endothelial marker 7-related protein).
TPBG	6: 83130067-83137264	TPBG	Trophoblast glycoprotein precursor (5T4 oncofetal trophoblast glycoprotein) (5T4 oncotrophoblast glycoprotein) (5T4 oncofetal antigen) (M6P1).
RET	10: 117812943-118022969	GFRA1	GDNF family receptor alpha-1 precursor (GFR-alpha-1) (GDNF receptor alpha) (GDNFR-alpha) (TGF-beta-related neurotrophic factor receptor 1) (RET ligand 1).
UBAP1	9: 34169003-34242521	UBAP1	Ubiquitin-associated protein 1 (UBAP).
FOS	14: 74815284-74818685	FOS	Proto-oncogene protein c-fos (Cellular oncogene fos) (G0/G1 switch regulatory protein 7).
COPS5	8: 68117871-68136785	COPS5	COP9 signalosome complex subunit 5 (EC 3.4.-.-) (Signalosome subunit 5) (SGN5) (Jun activation domain-binding protein 1).

Table 6: Colon cancer associated genes based on SNPs3D annotations. SNDs3D reference column describes the name of the gene associated to colon cancer. The corresponding genes of interesting regions are listed after the SNDs3D entry.

CCS ref.	locus	gene name	description
CSMD3	8: 113304337-114518418	CSMD3	CUB and sushi domain-containing protein 3 precursor (CUB and sushi multiple domains protein 3).
GNAS	20: 56848168-56919642	GNAS	Guanine nucleotide-binding protein G(s) subunit alpha (Adenylate cyclase-stimulating G alpha protein).
GUCY1A2	11: 106063120-106394381	GUCY1A2	Guanylate cyclase soluble subunit alpha-2 (EC 4.6.1.2) (GCS-alpha-2).
PKHD1	6: 51588057-52060382	PKHD1	Polycystic kidney and hepatic disease 1 precursor (Fibrocystin) (Polyductin) (Tigmin).
RET	10: 117812943-118022969	GFRA1	GDNF family receptor alpha-1 precursor (GFR-alpha-1) (GDNF receptor alpha) (GDNFR-alpha) (TGF-beta-related neurotrophic factor receptor 1) (RET ligand 1).
SCN3B	11: 123005107-123030165	SCN3B	Sodium channel beta-3 subunit precursor.

Table 7: Colon cancer associated genes based on [Sjö06] annotations. CCS reference column describes the name of the candidate cancer gene given in the article. The corresponding genes of interesting regions are listed after the reference name entry.

Chromosome	Start–End	Genes
1	98190565–98214180	
2	159328041–159328524	
3	100986418–101207563	*COL8A1*, *NP_997369.1*, *C3orf26*, *NP_878913.2*
4	165247274–165247492	
5	85666781–85667222	
5	86307720–86318185	
6	93576978–93577084	
6	95211104–95213139	
7	80525620–80526424	
7	116723355–116728976	*WNT2*
8	113348654–113368451	*CSMD3*
16	56409076–56409197	
21	37379327–37439134	*TTC3*

Table 8: Interesting deletion condidates and overlapping genes that were found in CRC study.

The analysis of `NoCall` SNPs revealed 13 candidate sites of possible homozygous deletions. The candidate sites consisted of at least two overlapping deletions in patient data, whereas no deletions regions were accepted in reference samples. A region is considered to be a deletion if it consisted of at least two reference SNPs or three SNPs in patient data. The physical loci of the interesting regions are shown in Table 8. Genes overlapping the interesting regions were fetched with RegionAnnotator and they have been listed on the right most column of the table.

8 Discussion

CohortComparator can be used as visualisation and comparison tool for two cohorts analysed with high density SNP-microarrays. The genome wide analysis is thought to require over 10^5 SNPs but the resolution differences in samples with different origin can be compensated using haplotype estimates of the population. The graphs of homozygous regions and their allelic constitutions can be used to inspect common patterns of haplotypes and the relationships between the samples. High rates of homozygosity may indicate familiar relatedness of individuals' parents. The relative amounts of homozygosity are readily shown in CohortComparator graphs, which help in estimating the distribution. The extensive size of the human genome leads to challenges in the visualisation of the data. CohortComparator takes an approach of splitting the data into chromosome specific graphs and highlighting the possibly interesting sample features based on their rareness in references. The clear separation between long homozygous regions and homozygous SNPs helps in concentrating into significant features of the data.

We found hidden Markov models unsuitable for detecting homozygous regions but they have been successfully used in many genetic studies. Therefore, CohortComparator provides a framework for HMMs than can be used to detect genomic regions defined by the two states of a given HMM. The framework provides a generic implementation of Viterbi algorithm than calculates the most probable state sequence for the given HMM and an emission sequence with the alphabet of the HMM. The transition probabilities can be given in a form of a function of previous and following states and an index of the emission sequence value. The generalisation of the HMM transition probability matrix allows the use of SNP specific transition probabilities, which may take the distances between the SNPs into account.

The selection of proper reference samples is a crucial step as the comparison between the data sets is extremely sensitive to references sharing the mutation in interest. The colorectal study was completed with a nil ($f.limit.r = 0$) tolerance of homozygous alleles, which means that a single reference sample with a latent susceptibility to CRC would cover the site of mutation. The tolerance limit can be loosened as more reference samples become available but some adjustments may be needed to work for low penetrance mutations unless a higher penetrance is archived in association to another mutation. The high penetrance combinations would be an interesting field of study but it requires a lot more samples.

Feasible results were archived using the sliding window method with parameters: $l = 55$,

$w = 20$, $g = 0.05$, and $s = 2$ (the parameters are given in SNPs instead of base pairs). The homozygous regions produced were long enough to be biologically feasible but the method was still capable of detecting short regions affecting single genes.

The list of possibly mutated genes produced by RegionAnnotator can be integrated with expression microarray data produced from the patient samples associated to the regions. The expression microarrays experiments are used to detect expression levels of various genes. The genes that are not expressed in their normal concentrations in patient samples are possibly inactivated by the inherited mutation. The comparison of RegionAnnotator genes and the list of inactive genes combine the genetic information to the actual functionality of the cells and separate the actively transcribed genes from the silent ones. The silent genes are not necessarily damaged by a mutation but the inactivation may be due the cellular state or a pure measurement error and the actively transcribed genes are not necessarily working as they should if a mutation has altered their constitution in a malign manner. Still, the list of genes detected by RegionAnnotator and underexpressed in expression studies is a good basis for further studies.

CohortComparator was applied to the colon cancer dataset and 2123 interesting regions were found that may harbour a cancer exposing allele. The gene annotation of these regions led to a list of 2257 possible genes of which 1643 genes were classified as known. Experimental verification of these genes is a laborious and time-consuming task. The sequencing of the candidate genes of associated CRC samples is in progress and the results of the verification will be published separately.

The 13 interesting sites of deletions suggested by the program were all confirmed to be false positives based on the fact that the sequences around each SNP were replicating in polymerase chain reaction experiments. The SNPs of the deletion sites expressed poor hybridisations even in those samples that did not fail for all three continuous SNPs. The observations suggest that the `NoCall` haplotypes are more likely to be technical artefacts than real overlapping deletions in both chromosomes. The false discoveries can be eliminated by combining the region information to copy number analysis [Zha04]. The copy numbers are not used within CohortComparator because that would have required the use of probe intensities instead of genotypes. The genotype based approach has been chosen to support various sources of data. The integration of the different modules, such as the copy number tools, is compatible with our plan to build components that can be combined together in a study specific manned.

The bioinformatics approach described in this thesis should be applicable to various studies of genetic origin of recessive phenotypes. CohortComparator will be publicly avail-

able[10] so that it can be tested and used in different studies. CSC has begun to optimise my code for their computer clusters. Our aim is to have CohortComparator preinstalled on CSC servers so that the Finnish biologists can use it together with other bioinformatic tools they are already familiar.

The current version of CohortComparator does not take population haplotypes into account when it compares homozygous regions between the cohorts. The SNP vice comparison of the regions may lead to fragments of interesting regions if a homozygous region in case cohort consists of SNPs of the rare allele with some SNPS of the common allele between them. Each sequence of common (sometimes caused by SNPs that do not vary between the samples) genotypes may interrupt the interesting region as the same allele becomes present in references. One common SNP genotype is capable of spliting an interesting region into two fragments, which consequently lowers the scoring of the region. We are planning to find better solutions to this challenge in near future.

[10]The official home page of CohortComparator is
`http://www.ltdk.helsinki.fi/sysbio/csb/downloads/CohortComparator/`.

References

Aal Aaltonen, L. e. a., Explaining the excess familial risk of colorectal cancer associated with mismatch repair deficient and stable tumours. unpublished draft, correspondence to Richard S Houlston, Tel:+44(0)2087224175, e-mail:Richard.Houlston@icr.ac.uk.

Aal98 Aaltonen, L. e. a., Incidence of hereditary nonpolyposis colorectal cancer and the feasibility of molecular screening for the disease. *N Engl J Med*, 338,21(1998), pages 1481–1487. URL http://content.nejm.org/cgi/content/abstract/338/21/1481.

Abe02 Abecasi, C. e. a., Merlin – rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics*, 30,1(2002), pages 97–101.

AD04 Armitage, P. and Doll, R., The age distribution of cancer and a multi-stage theory of carcinogenosis. *Int. J. Epidemiol.*, 33,6(2004), pages 1174–1179. URL http://ije.oxfordjournals.org.

Aff04 Affymetrix, Inc., *GeneChip mapping 100K assay manual*, 2004.

Aff06 Affymetrix, Affymetrix - GeneChip®Array Manufacturing, 2006. http://www.affymetrix.com/technology/manufacturing/index.affx. [1.9.2006].

Ame06 American Cancer Society, Cancer facts & figures 2006. Atlanta, US, 2006.

Ash00 Ashburner, M. e. a., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25,1(2000), pages 25–29.

Bau70 Baum, L. e. a., A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics*, 41,1(1970), pages 164–171.

Ber06 Beroukhim, R. e. a., Inferring loss-of-heterozygosity from unpaired tumors using high-density oligonucleotide snp arrays. *PLoS Computational Biology*, 2,5(2006), page e41. URL http://dx.doi.org/10.1371/journal.pcbi.0020041.

Bir06 Birney, E. e. a., Ensembl 2006. *Nucl. Acids Res.*, 34, pages D556–561. URL http://nar.oxfordjournals.org/cgi/content/abstract/34/suppl_1/D556.

Bra01 Brazma, A. e. a., Minimum information about a microarray experiment (miame)-toward standards for microarray data. *Nature Genetics*, 29,4(2001), pages 365–371. URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve\&db=pubmed\&dopt=Abstract\&list_uids=11726920.

Car06 Carlson, C. e. a., Direct detection of null alleles in SNP genotyping data. *Hum. Mol. Genet.*, 15,12(2006), pages 1931–1937. URL http://hmg.oxfordjournals.org/cgi/content/abstract/15/12/1931.

Chi06 Chiang, A. e. a., Homozygosity mapping with SNP arrays identifies TRIM32, an E3 ubiquitin ligase, as a Bardet-Biedl syndrome gene (BBS11). *PNAS*, 103,16(2006), pages 6287–6292. URL http://www.pnas.org/cgi/content/abstract/103/16/6287.

Cle79 Cleveland, W., Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74,368(1979), pages 829–836. URL http://links.jstor.org/sici?sici=0162-1459%28197912%2974%3A368%3C829%3ARLWRAS%3E2.0.CO%3B2-L.

Con06 Conrad, D. e. a., A high-resolution survey of deletion polymorphism in the human genome. *Nature Genetics*, 38,1(2006), pages 75–81. URL http://dx.doi.org/10.1038/ng1697.

Di05 Di, X. e. a., Dynamic model based algorithms for screening and genotyping over 100K SNPs on oligonucleotide microarrays. *Bioinformatics*, 21,9(2005), pages 1958–1963. URL http://bioinformatics.oxfordjournals.org/cgi/content/abstract/21/9/1958.

Fea98 Fearon, E. *The genetic basis of human cancer*, chapter Tumor suppressor genes, pages 229–236. McGraw-Hill, New York, 1998.

Fin05 Finnish Cancer Registry, Cancer in Finland 2002 and 2003. Helsinki, 2005.

Gar06 Garber, K., No miR hype: MicroRNAs cancer role expands. *Journal of the National Cancer Institute*, 98,13(2006), pages 885–887. URL http://www.ingentaconnect.com/content/oup/jnci/2006/00000098/00000013/art00885.

GT02 George, A. and Thompson, E., Multipoint linkage analysis for disease mapping in extended pedigrees: A Markov Chain Monte Carlo approach. Technical Report 405, Department of Statistics, University of Washington, Seattle, 2002.

HP04 Houlston, R. and Peto, J., The search for low-penetrance cancer susceptibility alleles. *Oncogene*, 23,38(2004).

HW00 Hanahan, D. and Weinberg, R., The hallmarks of cancer. *Cell*, 100,1(2000), pages 57–70.

Jaa03 Jaakson, K. e. a., Genotyping microarray (gene chip) for the ABCR (ABCA4) gene. *Human Mutation*, 22,5(2003), pages 395–403.

JH01 Johns, L. and Houlston, R., A systematic review and meta-analysis of familial colorectal cancer risk. *Am. J. Gastroenterol.*, 96,10(2001), pages 2992–3003.

KF00 Kam, P. and Ferch, N., Apoptosis: mechanisms and clinical implications. *Anaesthesia*, 55,11(2000), pages 1081–1093.

Knu71 Knudson, A., Mutation and cancer: Statistical study of retinoblastoma. *PNAS*, 68,4(1971), pages 820–823. URL http://www.pnas.org/cgi/content/abstract/68/4/820.

LB87 Lander, E. and Botstein, D., Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science*, 236,4808(1987), pages 1567–1570. URL http://www.sciencemag.org/cgi/content/abstract/236/4808/1567.

Lic00 Lichtenstein, P. e. a., Environmental and heritable factors in the causation of cancer–analyses of cohorts of twins from sweden, denmark, and finland. *N Engl J Med*, 343,2(2000), pages 78–85. URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve\&db=pubmed\&dopt=Abstract\&list_uids=10891514.

Lin04 Lin, M. e. a., dChipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data. *Bioinformatics*, 20,8(2004), pages 1233–1240. URL http://bioinformatics.oxfordjournals.org/cgi/content/abstract/20/8/1233.

LT00 Lindblad-Toh, K. e. a., Loss-of-heterozygosity analysis of small-cell lung carcinomas using single-nucleotide polymorphism arrays. *Nature Biotechnology*, 18,9(2000), pages 1001–1005.

Mid04 Middleton, F. e. a., Genomewide linkage analysis of bipolar disorder by use of a high-density single-nucleotide-polymorphism (snp) genotyping assay: a comparison with microsatellite marker assays and finding of significant linkage to chromosome 6q22. *Am J Hum Genet*, 74, pages 886–897.

MK06 Mizrachi-Koren, M. e. a., Homozygosity mapping as a screening tool for the molecular diagnosis of hereditary skin diseases in consanguineous populations. *Journal of the American Academy of Dermatology*, 55,3(2006), pages 393–401.

Moo05 Mooney, S., Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. *Briefings in Bioinformatics*, 6,1(2005), pages 44–56.

MYE98 Mueller, R., Young, I. and Emery, A., *Emery's Elements of Medical Genetics*. Churchill Livingtone, 10th edition, 1998.

Nan05 Nannya, Y. e. a., A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. *Cancer Res*, 65,14(2005), pages 6071–6079.

Nat06 National Human Genome Research Institute, genome.gov | National Human Genome Research Institute, 2006. http://www.genome.gov/. [14.9.2006].

Nic06 Nicolae, D. e. a., Coverage and charasteristics of the Affymetrix GeneChip Human Mapping 100K SNP set. *PLoS Genetics*, 2,5(2006), page e67.

Onk02 Onkamo, P. e. a., Association analysis for quantitative traits by data mining: QHPM. *Annals of Human Genetics*, 66,11(2002), pages 419–429.

PC06 Pounds, S. and Cheng, C., Robust estimation of the false discovery rate. *Bioinformatics*, page btl328. URL http://bioinformatics.oxfordjournals.org/cgi/content/abstract/btl328v1.

Pea06 Pearson, H., Genetics: What is a gene? *Nature*, 441,7092(2006), pages 398–401. URL http://dx.doi.org/10.1038/441398a.

PJV99 Peltonen, L., Jalanko, A. and Varilo, T., Molecular genetics of the finnish disease heritage. *Human Molecular Genetics*, 8,10(1999), pages 1913–1923. URL http://www.ingentaconnect.com/content/oup/hmg/1999/00000008/00000010/art01913.

Pre02 Press, W. e. a., *Numerical Recipes in C++: the art of scientific computing*. Cambridge university press, 2002.

R D06 R Development Core Team, *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006. URL http://www.R-project.org. ISBN 3-900051-07-0.

RS05 Rabbee, N. and Speed, T., A genotype calling algorithm for Affymetrix SNP arrays. *Bioinformatics*, 22,1(2005), pages 7–12. URL http://bioinformatics.oxfordjournals.org/cgi/content/abstract/22/1/7.

Sal00 Salovaara, R. e. a., Population-based molecular detection of hereditary nonpolyposis colorectal cancer. *J Clin Oncol*, 18,11(2000), pages 2193–2200. URL http://www.jco.org/cgi/content/abstract/18/11/2193.

SD03 Stephens, M. and Donnelly, P., A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Journal of Human Genetics*, 73,5(2003), pages 1162–1169.

Sjö06 Sjöblom, T. e. a., The consensus coding sequences of human breast and colorectal cancers. *Science*, page 1133427. URL http://www.sciencemag.org/cgi/content/abstract/1133427v1.

SS06 Scheet, P. and Stephens, M., A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet*, 78,4(2006), pages 629–644.

Ste43 Stern, C., The Hardy-Weinberg law. *Science*, 97,2510(1943), pages 137–188.

SW65 Shapiro, S. and Wilk, M., An analysis of variance test for normality (complete samples). *Biometrika*, 52,3-4(1965), pages 591–611. URL http://biomet.oxfordjournals.org.

Syv05 Syvänen, A. C., Toward genome-wide snp genotyping. *Nat Genet*, 37 Suppl. URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve\&db=pubmed\&dopt=Abstract\&list_uids=15920530.

The03 The International HapMap Consortium, The international HapMap project. *Nature*, 426,18(2003), pages 789–769.

Tin06 Ting, J. e. a., Analysis and visualization of chromosomal abnormalities in SNP data with SNPscan. *BMC Bioinformatics*, 7,1(2006), page 25. URL http://www.biomedcentral.com/1471-2105/7/25.

Vie06 Vierimaa, O. e. a., Pituitary adenoma predisposition caused by germline mutations in the AIP gene. *Science*, 312,5777(2006), pages 1228–1230. URL http://www.sciencemag.org/cgi/content/abstract/312/5777/1228.

Vit67 Viterbi, A., Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, pages 260–269.

VK02 Vogelstein, B. and Kinzler, K., *The genetic basis of human cancer*. McGraw-Hill, 2002.

VP04 Varilo, T. and Peltonen, L., Isolates and their potential use in complex gene mapping efforts. *Curr Opin Genet Dev*, pages 316–323.

Wan98 Wang, D. G. e. a., Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, 280,5366(1998), pages 1077–1082. URL http://www.sciencemag.org/cgi/content/abstract/280/5366/1077.

WCK06 Wagner, A., Creel, S. and Kalinowski, S. T., Estimating relatedness and relationships using microsatellite loci with null alleles. *Heredity*, 97,5(2006), pages 336–345.

Web06 Webb, E. e. a., Search for low penetrance alleles for colorectal cancer through a scan of 1467 non-synonymous SNPs in 2575 cases and 2707 controls with validation by kin-cohort analysis of 14 704 first-degree relatives. *Hum. Mol. Genet.*, 15,21(2006), pages 3263–3271. URL http://hmg.oxfordjournals.org/cgi/content/abstract/15/21/3263.

Wei06 Weinberg, R., *The Biology of Cancer*. Garland Science, Taylor & Francis Group, LLC, New York, US, 2006.

Woo04 Woods, C. e. a., A new method for autozygosity mapping using single nucleotide polymorphisms (SNPs) and EXCLUDEAR. *J Med Genet*, 41,8(2004), pages e101–. URL http://jmg.bmjjournals.com.

YMM06 Yue, P., Melamud, E. and Moult, J., SNPs3D: Candidate gene and SNP selection for association studies. *BMC Bioinformatics*, 7,1(2006), page 166. URL http://www.biomedcentral.com/1471-2105/7/166.

Zha04 Zhao, X. e. a., An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Res*, 64,9(2004), pages 3060–3071. URL http://cancerres.aacrjournals.org/cgi/content/abstract/64/9/3060.

Zha05 Zhao, X. e. a., Homozygous deletions and chromosome amplifications in human lung carcinomas revealed by single nucleotide polymorphism array analysis. *Cancer Res*, 65,13(2005), pages 5561–5570. URL http://cancerres.aacrjournals.org/cgi/content/abstract/65/13/5561.

ZPH05 Zheng, H., Peng, Z. and He, L., Loss of heterozygosity analyzed by single nucleotide polymorphism array in cancer. *World J. Gastroenterol.*, 11,43(2005), pages 6740–6744. URL http://www.wjgnet.com/1007-9327/11/6740.asp.

Appendix 1. Graphs of Interesting Regions

This appendix contains visualisations of SNPs in chromosome 5. The figures are generated using different parameter and data set combinations to illustrate some aspects of the analysis. The appendix can be used to get an idea of CohortComparator without a local installation and data sets.

CohortComparator can generate PostScript images of its results. Each chromosome is shown in its own image x-axis representing the nucleotide position from the first SNP to the last one. The samples are shown in rows so that the reference samples are at the top and the experimental data is at the bottom. Experimental samples are sorted based on the total length of heterozygous regions within the chromosome and the top most is the one with the greatest value.

Allele colours of the homozygous regions have been chosen so that red represents the allele present in homozygous regions found from the references or aa if the frequencies of aa and bb are equal. Blue allele is the one that is not so common in references and thus it is more interesting when present in experimental data. The blue regions are favoured over the red ones in their length so that they expand all the way to the first red SNP. Dark violet have been used as a compromise in those experimental regions that have no colour determining SNPs, which means that they are on a region without homozygous regions in reference set. Two dark violet regions may differ in their allelic constitution independently. Regions with possibly interesting differences have been emphasized with vertical blocks covering the regions. Homozygosity have been marked with yellow, deletions with green, and exceptional alleles with orange. CohortComparator has a score limit option that can be used to drop interesting regions below the given limit but the limit of 0 has been used for all graphs in this thesis. An optional detection of compound heterozygotes is used in Figure 21 where they have been shown in gray as overlapping interesting regions.

The following list explains the parameters and the data that are used in the graphs of this appendix.

- Figure 21 illustrates typical results with the parameters that were used in initial studies of homozygosity. Two tiny green lines that are almost invisible are over the horizontal black line separating the data sets. Positions of these lines are: 59741074–59810894 and 101298106–101298106 and they represent interesting regions with no compound heterozygotes on references.

- Figure 22 contains the same data as in Figure 21 but the missing values such as HindIII SNPs in reference samples and all `NoCall` SNP in all samples are estimated using fastPHASE haplotyping software. The same parameters were used for both data sets as they have the same resolution. The amount of short regions that were likely to be false discoveries has diminished in reference set leading to an expansion of the interesting regions.

- Figure 23 has been created with almost same parameters as Figure 21 but l and w were given in base pairs instead of SNPs. The limit l was estimated so that it was of the same fold with the shortest regions found in patient data of Figure 24.

- Figure 24 represents the comparison of chromosome 5 of the colon cancer patients to themselves. The reference set consists of XbaI chip data whereas the data set consists of both XbaI and HindIII measurements.

- Figure 25 is exactly the same as Figure 24 but now the reference consists of both XbaI and HindIII measurements and the data set consists of XbaI SNPs. The difference between these two images is huge suggesting that there are many false positives when the HindIII data is ignored.

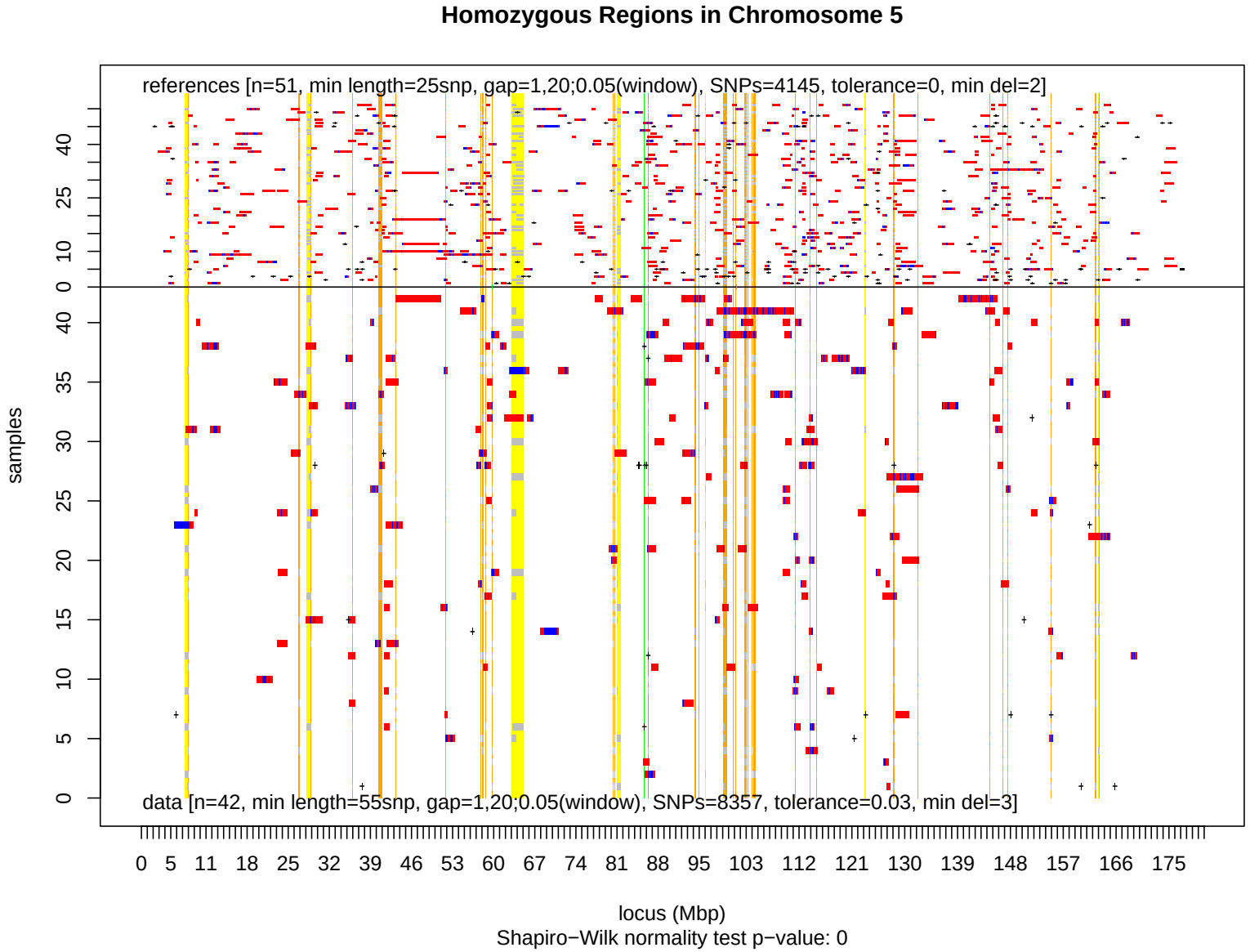

Figure 21: Typical set of interesting regions.

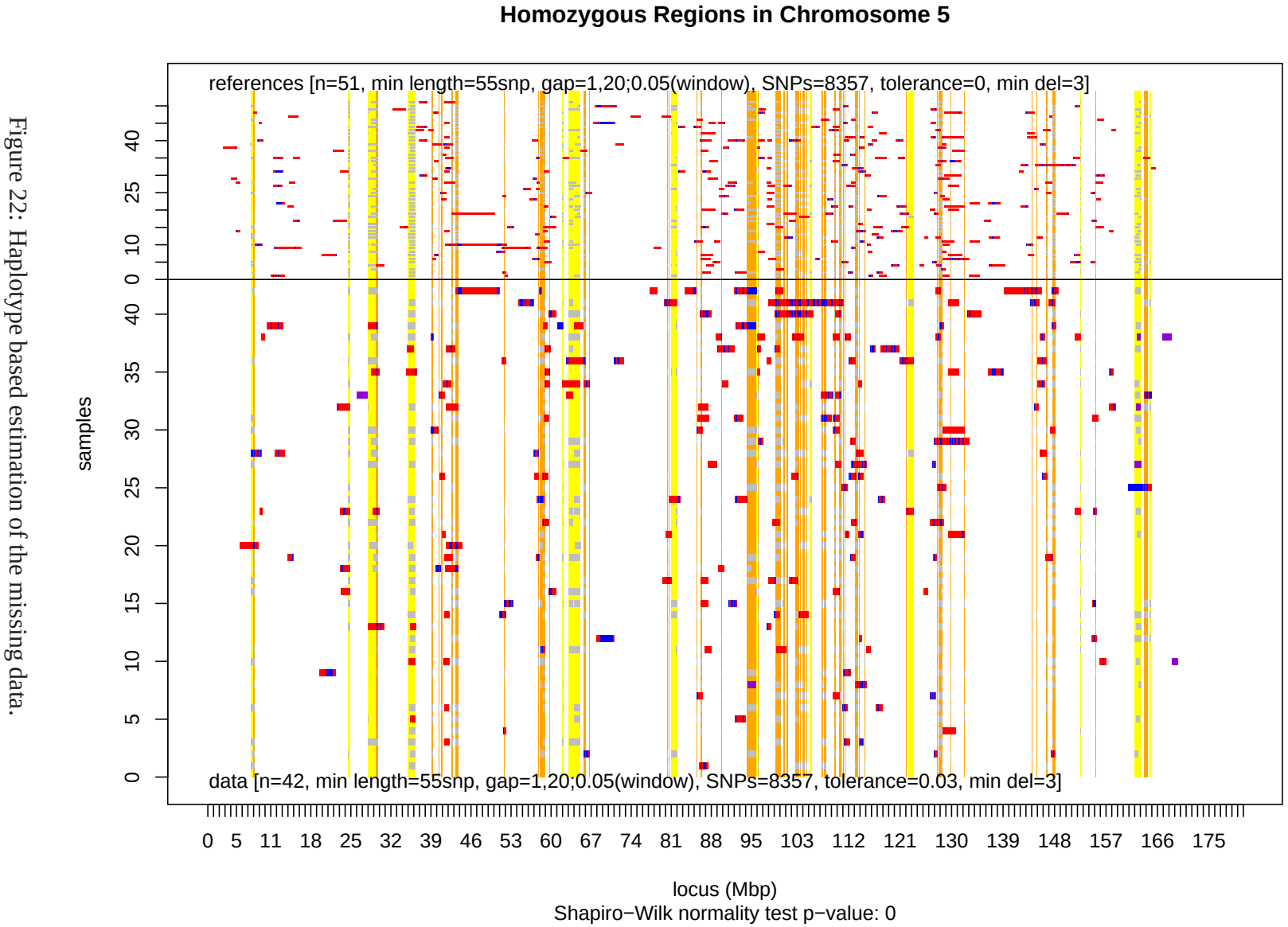

Figure 22: Haplotype based estimation of the missing data.

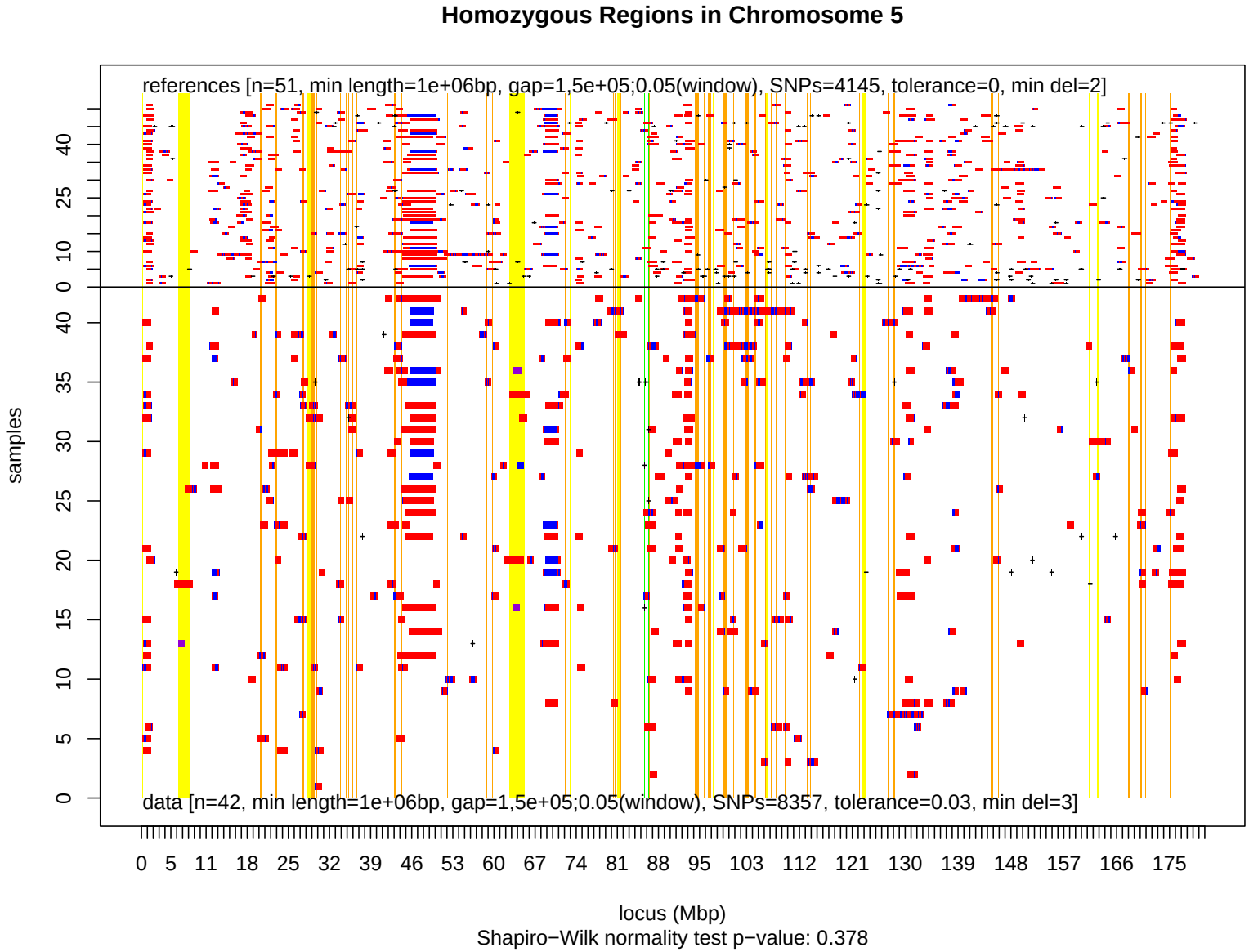

Figure 23: The use of bp metrics together with sliding window.

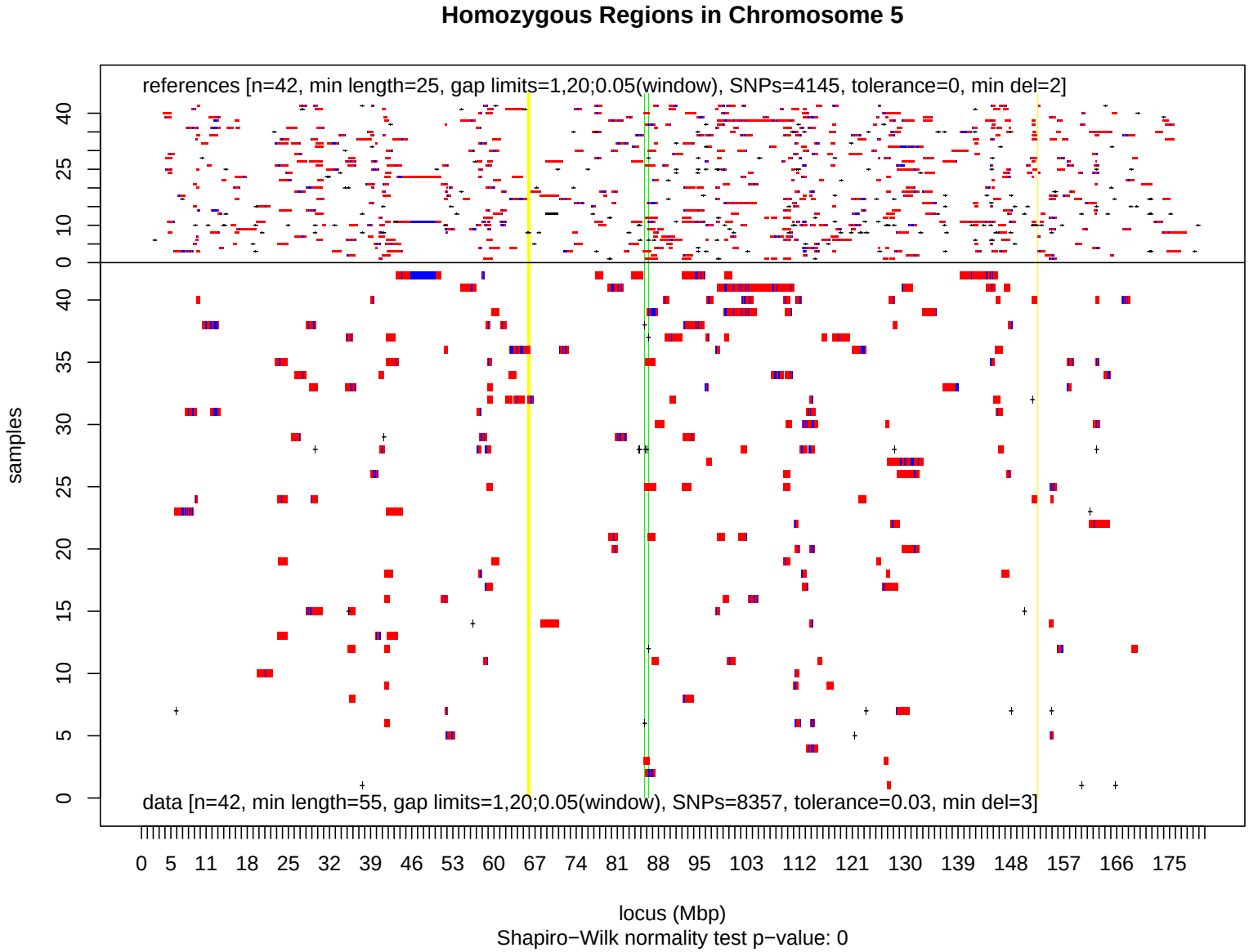

Figure 24: Self comparison of patient samples (reference=$\mathbf{S}^R_{Xbal}$, data=$\mathbf{S}^R$).

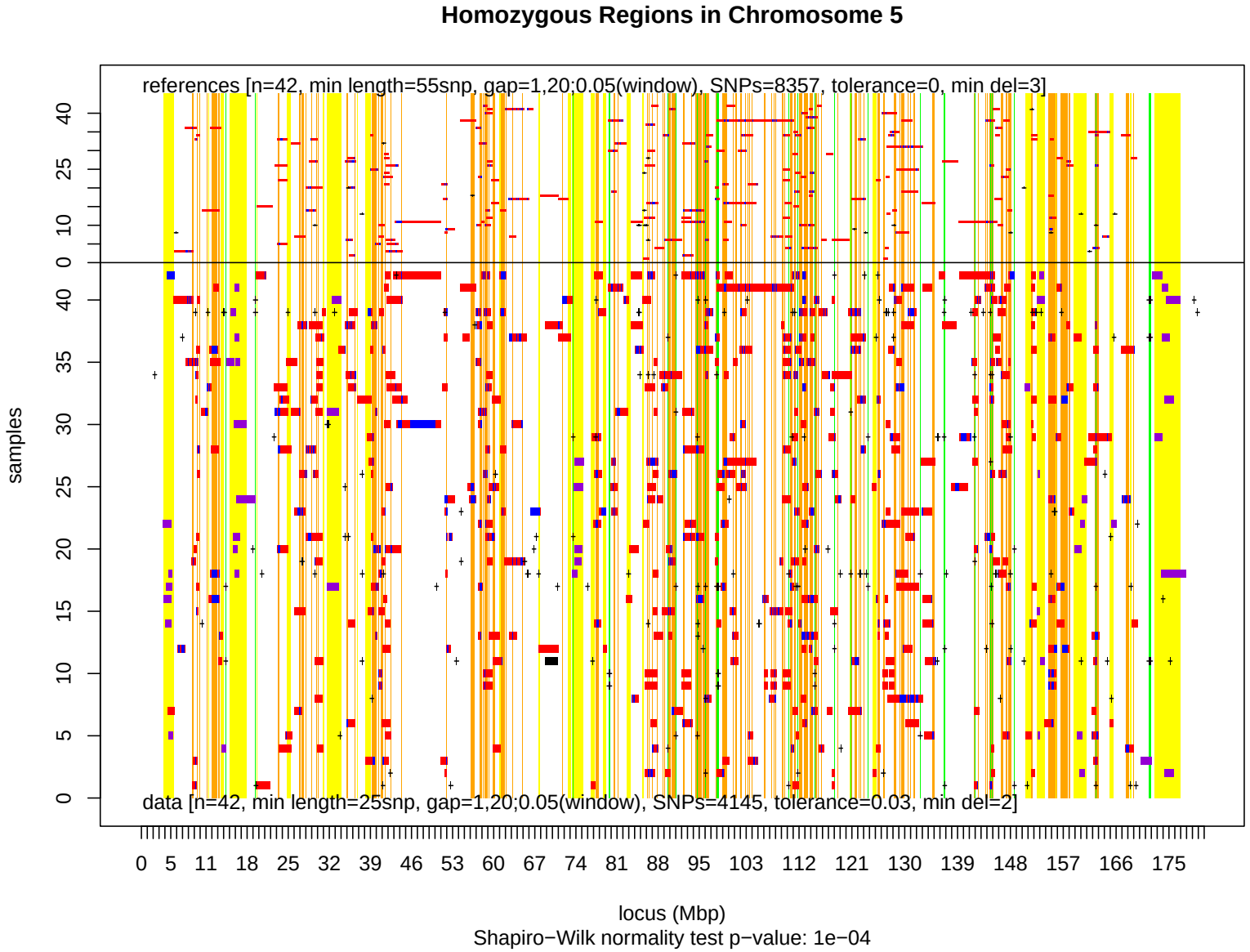

Figure 25: Self comparison of patient samples (reference=$\mathbf{S}^R$, data=$\mathbf{S}^R_{Xbal}$).

Appendix 2. Parameters of CohortComparator

All input parameters of CohortComparator are given in a separate R source file called `homozygous.properties`, which should be kept in the CohortComparator working directory. A short description of each parameter is given in here to bind the nomenclature of this document and the application.

`chromosome.range`	List of chromosome numbers for the genomic range of interest
`bpLength`	`TRUE` if l and w are given in base pairs and `FALSE` if they are given in SNPs
`d.limit.d`	Minimum number of continuous `NoCall` SNPs in data set considered to be a deletion
`d.limit.r`	Same as `d.limit.d` but for the reference set
`doCompound`	`TRUE` for activating the analysis of compound heterozygosity
`doInter`	This Boolean is used by the result file comparison tool. `TRUE` gives the intersection of regions in files `out.file.res.c1` and `out.file.res.c2` and `FALSE` substracts regions of `out.file.res.c2` out of `out.file.res.c1`.
`f.limit.d`	The minimum amount of overlapping features required in patient data for an interesting region (t^D) / n^D
`f.limit.r`	The maximum amount of overlaping features that can be ignored in reference set (t^R) / n^R
`gapRate.d`	g^D or (w^D, g^D) for sliding window
`gapRate.r`	g^R or (w^R, g^R) for sliding window
`HR_cfg_version`	Version number of CohortComparator compatible with the configuration file
`l.limit.d`	l^D
`l.limit.r`	l^R
`maxGap.d`	Maximal length of a hetorozygous section within a homozygous region in patient data (s^D)
`maxGap.r`	Maximal length of a hetorozygous section within a homozygous region in reference set (s^R)

out.file.results	File name for the list of interesting regions and related scores with sample names
out.file.res.c1	This Boolean is used by the result file comparison tool to give the file name of the first list of interesting regions.
out.file.res.c2	This Boolean is used by the result file comparison tool to give the file name of the second list of interesting regions.
out.folder	Directory name for the output files
region.def	Definition of homozygous region: hidden Markov model (hmm), gap distance (dist), gap ratio (ratio), or sliding window (window).
ref.height	Rendering parameter that tells the relative height of sample rows in references compared to those in data set
rID	An output file prefix for the files identifying the run
score.limit	An integer limit for the scores of interesting regions to be visualised in graphs. All regions will be shown if the limit is less than or equal to 0.
scoring.method	Either “fraction” or “total” depending of the chosen method of calculating scores of interesting regions as explained in Section 4.3
src.folder	Directory name for the input files
src.prefix.data	File name prefix for the data set SNP files
src.prefix.locD	File name prefix for the chromosome specific list of physical locations of SNPs in data set haplotype matrix
src.prefix.locR	File name prefix for the chromosome specific list of physical locations of SNPs in reference set haplotype matrix
src.prefix.refs	File name prefix for the reference set SNP files